COLLECTION-LEVEL SUBJECT ACCESS IN AGGREGATIONS OF DIGITAL COLLECTIONS: METADATA APPLICATION AND USE

BY

OKSANA LVIVNA ZAVALINA

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Professor Carole L. Palmer, Chair
Assistant Professor Kathryn La Barre
Associate Professor Allen Renear
Professor Dietmar Wolfram, University of Wisconsin (Milwaukee)

Abstract

Problems in subject access to information organization systems have been under investigation for a long time. Focusing on item-level information discovery and access, researchers have identified a range of subject access problems, including quality and application of metadata, as well as the complexity of user knowledge required for successful subject exploration. While aggregations of digital collections built in the United States and abroad generate collection-level metadata of various levels of granularity and richness, no research has yet focused on the role of collection-level metadata in user interaction with these aggregations. This dissertation research sought to bridge this gap by answering the question "***How does collection-level metadata mediate scholarly subject access to aggregated digital collections?"*** This goal was achieved using three research methods:

- in-depth comparative content analysis of collection-level metadata in three large-scale aggregations of cultural heritage digital collections: Opening History, American Memory, and The European Library
- transaction log analysis of user interactions, with Opening History, and
- interview and observation data on academic historians interacting with two aggregations: Opening History and American Memory.

It was found that subject-based resource discovery is significantly influenced by collection-level metadata richness. The richness includes such components as:

1) describing collection's subject matter with mutually-complementary values in different metadata fields, and

2) a variety of collection properties/characteristics encoded in the free-text *Description* field, including types and genres of objects in a digital collection, as well as topical, geographic and temporal coverage are the most consistently represented collection characteristics in free-text *Description* fields.

Analysis of user interactions with aggregations of digital collections yields a number of interesting findings. Item-level user interactions were found to occur more often than collection-level interactions. Collection browse is initiated more often than search, while subject browse (topical and geographic) is used most often. Majority of collection search queries fall within FRBR Group 3 categories: *object*, *concept,* and *place*. Significantly more *object*, *concept*, and *corporate body* searches and less *individual person*, *event* and *class of persons* searches were observed in collection searches than in item searches. While collection search is most often satisfied by *Description* and/or *Subjects* collection metadata fields, it would not retrieve a significant proportion of collection records without controlled-vocabulary subject metadata (*Temporal Coverage*, *Geographic Coverage*, *Subjects,* and *Objects*), and free-text metadata (the *Description* field). Observation data shows that collection metadata records in Opening History and American Memory aggregations are often viewed. Transaction log data show a high level of engagement with collection metadata records in Opening History, with the total page views for collections more than 4 times greater than item page views. Scholars observed viewing collection records valued descriptive information on provenance, collection size, types of objects, subjects, geographic coverage, and temporal coverage information. They also considered the structured display of collection metadata in Opening History more useful than the alternative approach taken by other aggregations, such as American Memory, which displays only the free-text *Description* field to the end-user.

The results extend the understanding of the value of collection-level subject metadata, particularly free-text metadata, for the scholarly users of aggregations of digital collections. The analysis of the collection metadata created by three large-scale aggregations provides a better understanding of collection-level metadata application patterns and suggests best practices. This dissertation is also the first empirical research contribution to test the FRBR model as a conceptual and analytic framework for studying collection-level subject access.

Acknowledgements

I am very grateful to my dissertation advisor and chair of research, Professor Carole Palmer, for inspiring me, helping to shape my research agenda, and guiding me through this dissertation journey.

Special thanks to my dissertation committee members, professors Kathryn La Barre, Allen Renear, and Dietmar Wolfram, for their thoughtful comments and advise throughout the whole process, from the first draft of the dissertation proposal to the completed dissertation thesis.

I would like to thank the IMLS Digital Collections and Content project team at the University of Illinois, the current and former members of which have been a valuable source of information on Opening History aggregation for my research. I am also grateful to Sally Chambers (The European Library interoperability manager) and Christa Maher (American Memory metadata librarian), who provided me with the data on their respective aggregations of digital collections, without which this research would be incomplete.

I am grateful to the members of the Metadata Roundtable for their thoughtful review of my coding and to the members of GSLIS Students Research Group for participation in intercoder reliability tests.

Special thanks to my friends and colleagues Jan Adamczyk, Susanne Birgerson, Ellen Knutson, and Robert Howerton for their great help with proofreading various drafts of this dissertation thesis.

I am also very thankful to my family and friends for their unconditional love, patience, and support.

Table of Contents

List of Figures

List of Tables

Chapter 1. Background of this Study

1.1 Introduction

Problems in subject access to information organization systems — specifically catalogs and bibliographic databases — have long been investigated in the field of library and information science (e.g., Jackson, 1958; Lipetz, 1970; Tagliacozzo & Kochen, 1970; Matthews, Lawrence, & Ferguson, 1983). Focusing on item-level information discovery and access, researchers have identified a range of subject access problems. These include limitations within the systems, such as those related to quality and application of controlled vocabularies (e.g., Cochrane, 1986, 2000), and a range of issues related to user interactions with these systems. For example, user experiences are influenced by their subject domain and conceptual knowledge (e.g., Allen, 1991; Borgman, 1996; Markey, 2007), their understanding of how the information organization system functions (Borgman, 1996), and their knowledge of sources to search and ordering of their searches (Markey, 2007).

In the past decade, considerable work has gone into building aggregations of digital collections in the United States (e.g., American Memory, Opening History, etc.) and in Europe (e.g., The European Library). While some of these aggregations (e.g., OAIster, Europeana, etc.) do not use collection-level descriptions, others consider collection metadata important for providing context for the digital items harvested from distributed collections. In building these resources, collection level metadata records are created to describe digital collections in an aggregation, and these records have expressed collection-level descriptions at various levels of granularity and richness. With a number of collection-level metadata systems now in place, it is an opportune time to analyze these practices to provide a better understanding of the roles and

functions of collection-level metadata in large digital aggregations. While there is considerable research on defining and describing digital collections from the digital resource developers' perspective (e.g., Lee, 2000; Manoff, 2000; Currall, Moss, & Stuart, 2004; Palmer et al., 2006), no research has yet focused on the role of collection-level subject metadata in the context of user interactions with aggregations of digital collections, nor on user strategies in dealing with collection-level information discovery.

This dissertation research bridges this gap by seeking answers to the following question: How does collection-level metadata mediate scholarly subject access to aggregated digital collections? More specifically, three aggregations of digital collections are examined in terms of how subjects are represented in collection-level metadata, but also from the perspective of scholarly users in the humanities and the social sciences—specifically how they interact with aggregations of digital items rather than with individual items, and what role collection-level metadata (free-text and structured) richness plays in such interactions. This study will extend the field's understanding of the role and the perceived value of collection-level subject metadata, particularly free-text metadata, for this target user community.

This chapter reviews the relevant research to provide the appropriate context for the study. It covers important literature on subject access and on conceptual models of subject representation: *Functional Requirements for Bibliographic Records, Functional Requirements for Authority Data,* and *Functional Requirements for Subject Authority Records*. This review includes:

1. catalog use studies that focus on subject access and user interaction with the information organization systems are included,

2. information seeking behavior research focusing on users' knowledge of, experience with, and expectations for digital libraries,
3. key publications on digital collections and collection-level metadata.

Figure 1 located at the end of this chapter (section 1.7) illustrates the position of this dissertation research among these related areas within the field of Library and Information Science. A glossary of the terms applied in this dissertation is included in Appendix F.

1.2 Subject Access

Although a universally accepted definition of *subject* or *subject matter* has never been developed in Library and Information Science (LIS), subject access has been one of the central topics for decades, particularly in regard to information seeking and information retrieval (IR) theory (Hjørland, 1997). As defined by Cochrane (1979), *subject access* means systematic (e.g., classification system), topical (e.g., subject headings), and natural (e.g., title, abstract words) approaches to the subject matter in a collection and encompasses both processes of "subject cataloging and retrieval by the searcher."

Subject access in catalogs and other information organization systems is provided through metadata. Metadata is commonly defined simply as "data about data," but for online catalogs and similar digital resources, subject metadata is generally provided within structured records that describe information objects or collections of objects. Hereafter, the term "metadata" will be used as defined by the *Encyclopedia of Library and Information Science* (Greenberg, 2005) — "structured data about an object that supports functions associated with the designated object."

Charles Ammi Cutter recognized subject access as a fundamental function of the library catalog in 1876. He formulated the three major objectives of the library catalog as:

1. To enable a person to find a book of which either the author, the title, or the **subject** is known,
2. To show what the library has by a given author, in a given **subject**, in a given kind of literature,
3. To assist in choice of a book as to its edition (bibliographically), as to its character (literary or **topical**).

Cutter's principles have been accepted for over a century as the framework for defining the basic tasks with which catalogs should assist library users. However, the Paris Principles—a set of cataloging principles formally adopted by the International Federation of Library Associations (IFLA) in 1961, began with Cutter's principles and were further revised by Lubetzky (1960)—focusing on descriptive cataloging. Subject access issues were ignored or underrepresented in statements of cataloging principles. Subject access was not accounted for until IFLA's working group on Functional Requirements for Bibliographic Records (IFLA, 1997; 2008) included subject as a relation and defined a set of subject entities in its FRBR entity-relationship model of the bibliographic universe.

1.2.1 Functional Requirements for Bibliographic Records and Related Models of Subject Representation

The Functional Requirements for Bibliographic Records (FRBR) entity-relationship model of the bibliographic universe (FRBR, 1997, 2008) has been increasingly influential in thinking about the ways to organize descriptive metadata in databases. The four user tasks

defined by FRBR as *find, identify, select,* and *obtain*, include the following subtasks in extension of Cutter's principles and specifically designed to deal with subject access:

1. find the works on a given subject,
2. find the works in which a concept is significantly treated,
3. select a work by its main subject only,
4. search for works on related subjects,
5. search for works in which related or connected subjects are handled.

Based on the results of user studies, the Functional Requirements for Subject Authority Records (FRSAR) IFLA working group has recently adapted and refined the list of user tasks proposed by FRBR to reflect the subject-specific user needs:

Find: To find a subject entity or set of entities corresponding to stated criteria.

Identify: To identify a subject entity based on certain attributes/characteristics.

Select: To select a subject entity.

Obtain: To obtain additional information about the subject entity and/or to obtain bibliographic records or resources about this subject entity.

Explore: To explore relationships between subject entities, correlations to other subject vocabularies and structure of a subject domain (Žumer, Salaba, & Zeng, 2007).

Figure 2 in section 1.7 of this thesis illustrates the entities and subject relationships in the FRBR model. While it is acknowledged by FRBR model developers that each of the ten entities in the model (*work, expression, manifestation, item, person, corporate body, concept, object, event,* and *place*) can serve as a subject of a work, FRBR's Group 3 entities includes *concept, object, event,* and *place* — the major types of a subject:

- *"concept*: an abstract notion or idea, encompasses a comprehensive range of abstractions that may be the subject of a *work*: fields of knowledge, disciplines, schools of thought (philosophies, religions, political ideologies, etc.), theories, processes, techniques, practices, etc. A *concept* may be broad in nature or narrowly defined and precise." (FRBR 2008, p. 26)
- *"object*: a material thing, encompasses a comprehensive range of material things that may be the subject of a *work*: animate and inanimate objects occurring in nature, fixed, movable, and moving objects that are the product of human creation, objects that no longer exist." (p. 27)
- *"event*: an action or occurrence, encompasses a comprehensive range of actions and occurrences that may be the subject of a work: historical events, epochs, periods of time, etc."(p. 28)
- "*place*: a location, encompasses a comprehensive range of locations: terrestrial and extra-terrestrial, historical and contemporary, geographic features and geo-political jurisdictions."(p. 28-29).

While Group 1[1] and Group 2[2] entities in the FRBR model are well defined and each is accompanied by a list of attributes and characteristics, Group 3 entities have limited definitions and lack elaborated characteristics. Both the initial (1997) and updated (2008) versions of the FRBR model list only one attribute for each of the Group 3 entities — the term for the entity — with two characteristics under it: a subject heading and a classification number. However, the Functional Requirements to Authority Data conceptual model (FRAD, 2007), released by the Functional Requirements and Numbering of Authority Records (FRANAR) IFLA working

[1] *Work, expression, manifestation,* and *item.*
[2] *Person* and *corporate body*

group, which is closely related to the FRBR Review Group, expanded the list of attributes for most of the FRBR entities. In the FRAD model, the *concept* entity has a newly-added "type of concept" attribute. The *object* entity received 5 new attributes: "type of object", "date of production", "place of production", "producer/fabricator," and "physical medium." The list of *event* attributes was expanded to include "date associated with the event" and "place associated with the event," while "coordinates" and "other geographical information" were added to the attributes of the *place* entity.

The set of FRBR's ten entities is not intended to be comprehensive and is likely to be expanded. Currently, the IFLA working group on Functional Requirements for Subject Authority Records (FRSAR) is working to define subjects, by focusing on the Group 3 entities. For example, researchers (e.g., Zeng & Salaba, 2005; Delsey, 2005) have suggested revisiting Group 3 by adding *time* and *process*, differentiating between a dynamic *event* and a static *situation*, between *concrete* and *abstract concepts*. More recently, Maxwell (2008) has pointed out that *genre/form* could be considered a subclass of the *concept* entity. The FRSAR working group has also suggested that Group 2 FRBR entities, currently including *person* and *corporate body*, should incorporate a third entity — *family* (Zeng & Salaba, 2005). Unlike FRBR, the FRAD conceptual model includes the *family* entity while also modifying the definition of the *person* entity by including groups of people working under the same pseudonym (e.g., Ellery Queen) or trademark (e.g., Betty Crocker). As pointed out by Maxwell (2008), relationships between Group 2 entities (*person*-to-*person*, *person*-to-*persona*[3], *person*-to-*corporate-body*), and between a *person* entity and Group 3 entities (e.g., "Edit Piaf" and "Singers", "Actresses", "Authors, French") have also been overlooked by the FRBR model. Moreover, the FRBR model has been

[3] "Persona established or adopted by individual or group" (FRAD, p.8) refers to individual or joint pseudonyms.

criticized for a lack of granularity relating to groups of individuals other then *corporate body* (e.g., communities, societies, etc.), as they are lumped together without differentiation under the *object* or *concept* entities (Delsey, 2005). At least two supersets of individual persons, which appear to be used in actual searches — *ethnic group* and *class of persons* (Zavalina, 2007) — are not currently accommodated by the FRBR model. Another important area in need of augmentation is the omission of *collection of works* by the FRBR model. Some of FRBR's adaptations for describing specific kinds of materials attempt to alleviate this and other problems. For example, the ePrints Application Profile entity-relationship model (*ePrints* …, 2007) lists "collection" as a possible value for the "is part of" attribute of *copy*[4].

While introducing much complexity, none of the evolving models of the "bibliographic universe" seems adequate to cover subject access. They provide an incomplete picture of functions, mostly because virtually no evidence on the real-world functions and use has been incorporated, especially in the realm of subject searching. Results of a recent survey (Zhang & Salaba, 2007a) demonstrate that the "need to verify and validate the FRBR model against real data and in different communities to make sure the model is valid and applicable" is among top 10 critical issues and challenges within the FRBR research and development community. Meeting this challenge is impossible without research into how subject searching is done in practice, and this urgent need has been emphasized by the members of IFLA FRBR review working group (Riesthuis & Žumer, 2004).

One of the aims of this dissertation research was to analyze subject-specific user interactions with an information organization system in a real-life situation with one user community (academic historians using an aggregation of cultural heritage digital collections).

[4] *Copy* in ePrints model corresponds to FRBR *item*.

The results have been compared with the entities of the FRBR and related (FRANAR and FRSAR) conceptual models. The next section will summarize the findings of catalog use studies, which provide necessary historical context for such an exploration: the patterns of user interaction with metadata of print and online catalogs, and the ways in which this metadata has been found to facilitate or hinder subject access to library collections.

1.3 Subject Access in Catalog Use Studies

According to Lee (2003), "purposive exploration on chosen subjects" (including subject search and subject browse) is one of the three general types of scholarly information seeking; the two others are "locating specific information and/or documents," and "general scanning for nonspecific information." Catalog and bibliographic database use studies have revealed a lot of information about this "purposive exploration" and subject access in general, which will be briefly overviewed in this section.

As defined by Lipetz (1970), subject search is a "search, where the user seeks to identify publications on a known abstract topic." Historically, subject search has been recognized as one of the two major approaches employed by users in catalog searching, together with the known-item — i.e. author/title — approach (Krikelas, 1972). Subject search is sometimes further subdivided into unknown search and area search (Slone, 2000).

Catalog use studies have demonstrated that users initiate subject searching in catalogs more often in public than academic libraries. Undergraduate students and other beginners in a scholarly field have been noted to use subject search more than graduate students, faculty, and other experts (Jackson, 1958; R. Palmer, 1970; Tagliacozzo & Kochen, 1970; Lipetz, 1970; Larson, 1991a). In studies of card catalog and early online catalog use, subject search was

generally found to be less used overall than known-item search, though research results vary substantially: subject search was used by 44-57% of searches in the 1950s ALA Catalog Use Study depending on the type of library and the group of searchers (Jackson, 1958), 21-46% in the late 1960s (R. Palmer, 1970; Tagliacozzo & Kochen, 1970; Lipetz, 1970), 59% in the early 1980s (Matthews, Lawrence, & Ferguson, 1983), 76% in the mid 1980s (Larson, 1991b) and 40-46% in the late 1980s (Peters, 1991; Larson, 1991b).

Such a variation in subject search use can be explained to a large extent by the practical difficulty of distinguishing subject searches from other kinds of searches. For example, in interviewing catalog users, Lipetz (1970) made an interesting observation that although in reality the majority of searches were performed with the goal of retrieving information on a particular topic, users generally reported looking for specific documents. He suggested that the users' search behavior is shaped by the available features of an information organization system, and that if library catalogs were better suited for subject searching, there would be less overt known-item searching and more explicit subject searching. Two decades later, Peters (1991) reported similar observations in his transaction log analysis: remote users of online public access catalogs seemed to favor title keyword search as a type of subject searching, while the vast majority of in-house users were using subject browsing rather than subject searching with controlled-vocabulary terms (LCSH).

Negative user experiences (Krikelas, 1972; Lipetz, 1970; Markey, 1984), particularly search failure and information overload in subject search (Larson, 1991a), have been identified as the major reason for this apparent underuse of controlled vocabulary searches. Improvements made to the underlying search structure of the catalog (Library of Congress Subject Headings, Library of Congress and Dewey Decimal classifications), the shortcomings of which contribute

greatly to this negative experience, remain insufficient (e.g., Cochrane, 2000). For example, the lack of specificity and exhaustivity of LCSH vocabulary, inconsistency in heading structure and syndetic structure[5], outdatedness and bias in many subject terms, lack of notes and links between LCC classification numbers and LCSH (or other controlled vocabularies) have been affecting subject search performance in catalogs. The structure of subject headings (i.e., postcoordination versus precoordination) has often been named as one of the factors complicating users' experience in the subject searching of card and online catalogs (e.g., Taube, 1953; Farradine, 1970; Weinberg, 1995). Providing a postcoordinate (faceted) approach to controlled-vocabulary subject metadata became even more desirable goal in the Web environment, (e.g., Chan & Hodges, 2000).

The Council on Library Research nationwide catalog use survey (Matthews, Lawrence, & Ferguson, 1983) recommended introducing the keyword search option to compensate for the complexity of controlled-vocabulary subject search. Since then, keyword search has been actively used as a variety of subject search that does not require controlled vocabulary. Other CLR study recommendations included increasing the amount of subject information in bibliographic records, permitting users to browse the subject index or thesaurus, and restricting the number of possible search terms "either by rigorously controlling the vocabulary or by automatically linking the user's search terms with synonymous and related terms that appear in subject headings" (p. 178-179).

[5] According to Birger Hjørland's definition, syndetic structure is the system of "see" and "see also" cross references to other indexing terms in catalog, which is used to connect related headings by means of cross-references (http://www.db.dk/bh/lifeboat_ko/CONCEPTS/syndetic_structure.htm)

Although the multiplicity of vocabularies involved in search query processing — including those of authors, documents, searchers, indexers, syndetic structures, and queries — make the mismatch between them an ever-existing possibility (Buckland, 1999), the following recommendations proved valuable in enhancing subject access and improving user subject search experience in online catalogs:

- using post-Boolean probabilistic searching with automatic spelling correction, term weighting, intelligent stemming, relevance feedback, and output ranking (Hildreth, 1989; 1995; Drabenstott, 1991),
- adding tables of contents and back-of-the-book indexes to bibliographic records (Atherton, 1978; Wormell, 1981; Markey & Calhoun, 1987),
- expanding the online catalog with full text (Drabenstott, 1991),
- increasing finding strategies in online catalogs through the library classification (Markey & Demeyer, 1986; Larson, 1991c).

As can be seen from the summary above, the impact of the richness of library catalog metadata on the search performance has been studied by a number of researchers. According to Krikelas (1972), one of the fundamental questions on many catalog use studies was "Does the amount of bibliographic information affect the utility of the catalog?" Different bibliographic description elements (fields) in traditional card catalogs have seen varying degrees of use among library patrons. A number of catalog use studies, summarized by Krikelas (1972), reported heavy use of author, title, subject headings, call number, and date of publication, while place of publication, publisher, edition, and content note tended to be consulted less often, and size, series note, and illustration statement were rarely used by library patrons. This even led some researchers in early the 1970s (e.g., Palmer, 1970) to conclude that minimizing bibliographic

information in catalog entries for use in the first generation electronic catalogs would not decrease the success rate of user searching. Although the Machine Readable Catalog Record (MARC) standard developed for the use in online catalogs since the late 1960s provides a structure for rich encoding and content designation, this richness is not necessarily included in bibliographic records (e.g., Moen & Benardino, 2003).

1.4 User Knowledge, Expectations and Experiences

Largely because online catalogs brought in new affordances of "search capabilities indexing"[6] (Bates, 1989), users initially expressed much greater satisfaction with online catalogs than with traditional library card catalogs (e.g., Matthews, Lawrence, & Ferguson, 1983; Larson, 1991a). However, it has been later noted by numerous studies of online catalogs that most users find OPACs "disappointing, frustrating, illogical, counter-intuitive, and intimidating," which outweighs users' "appreciation, even admiration, of the 'control and order' of library-style environments" (Bawden & Vilar, 2006). Subject access has been found to be the most problematic both in card and online catalogs (Larson, 1991a).

1.4.1 Domain Knowledge and Subject Access

Borgman (1986) pointed out that failure to incorporate sufficient understanding of searching behavior, knowledge, and skills of users in design makes bibliographic databases in general (and their subject access capabilities in particular) hard to use. Conceptual knowledge was identified as one of the three layers of knowledge the user needs to perform searching in different databases (Borgman, 1996). The majority of user search problems were found to occur at this layer, which includes understanding the contents of database, knowing when to use which access point, using

[6] According to Bates (1989), online search capabilities include keyword searching, Boolean searching, truncation, and multi-index searching.

alternative search paths, formulating searches, understanding relations between different topics within a discipline, understanding and using keyword searching, distinguishing between no matches due to a search error and due to item unavailability in the database, and ways to narrow and broaden search results. Domain expertise, which is an important component of conceptual knowledge, was identified as one of the most important factors in use of online catalogs (Markey, 2007). Information seeking behavior and the outcomes of the search depend to a large extent on a searcher's level of knowledge both on a specific search topic and the broader subject domain (e.g., Allen, 1991).

Empirical research results emphasize the role of user domain knowledge in subject access and the effect it has on information seeking behavior. Studies reveal that both selection of the search terms and overall search tactics change with increase of domain knowledge. Wider and more specific/unique vocabulary is used in subject search by users with higher domain knowledge (Pennanen, Serola, & Vakkari, 2003; Hembroke et al., 2005). The query length (the number of searches per session) and search term reformulation behavior decrease, while the use of advanced search options and the query length (number of words per query) increase (Wildemuth, 2003; Zhang, Anghelescu, & Yuan, 2005; Hembrooke et al., 2005). Domain experts were found to have clear expectations for both the answer to the search question and the context in which it would appear (Marchionini et al., 1993). Scholarly users of the library databases preferred known-item searches or applied more self-constructed terms in the subject search within their domain, while utilizing controlled-vocabulary search terms and synonyms for subject search outside their domain (Hsieh-Yee, 1993; Connaway, Johnson, & Searing, 1997). Subject headings were usually too broad to pinpoint domain expert's specialized research interests (Connaway, Johnson, & Searing, 1997). The information seeking behavior studies

found that library catalogs were not designed to take advantage of "subject expert's knowledge" (Bates, 1972).

Like metadata, discussed in Section 1.2, domain knowledge has been found to be an important factor in the item-level subject access of scholarly users. This dissertation research was based on the logical assumption that domain knowledge also plays an important role in collection-level scholarly subject access. The two following sections analyze another important component of subject access: expectations of users (scholarly historians in particular) towards modern information organization systems, including both bibliographic databases and full-text digital libraries.

1.4.2 Expectations and Experiences

In the 1980s and 1990s, research suggested that in the long-run, online catalogs should be judged by their success in *answering questions* rather than *matching queries,* that more attention should be paid to supporting browsing as a prevailing form of search, and that exploratory design models are needed (Borgman, 1996; Hildreth, 1995). In the late 1990s and 2000s, user experiences using online catalogs, article databases, and digital libraries, significantly differ from experiences with other modern resource discovery tools, and the latter seems to influence the former. For example, user expectations of digital library services (Bawden & Vilar, 2006) are shaped by user experiences with major search engines (predominantly Google) which are familiar and easy to use,[7] widely used transactional sites (e.g., Amazon and eBay), popularity of computer games, and changes in the Western society in general (i.e., greater speed of developments, perceived need for immediate gratification, more information rich environment,

[7] Related ease of use is supported by research results showing that improvement in searching skills brings better results from library databases but not from Internet search engines (e.g., Brophy & Bawden, 2005).

and the popular heuristic of "satisficing"). As a result, users typically expect much more from digital libraries than from conventional library services. These expectations include: comprehensiveness, accessibility, immediate gratification, followability of data, ease of use, and multiple formats (Bawden & Vilar, 2006). Expectations of digital library services are often too high (although this is somewhat context-dependent), and are combined with a surprising lack of appreciation of basic points, such as that digital library collections being created based on the knowledge of user groups' needs. Bawden and Vilar also point to the fact that user expectations, including "collection expectation"[8] differ by user domain and level of expertise.

Research indicates user preference for using the Web[9] to search for information over online catalogs and article databases (e.g., Becker, 2003; Fast & Campbell, 2004; OCLC, 2006; Griffith & Brophy, 2005). Researchers routinely observe users following the way of least effort and selecting much less effective but very simple modes of searching, and the preference for very simple searching plays an important role in the selection of the search engines as the first choice (e.g., Griffith & Brophy, 2005). A study recently released by British Library researchers (*Information Behaviour*, 2008) found that the main characteristics of user behavior in virtual libraries include: horizontal information seeking (a form of skimming activity, where searchers view just one or two pages from an academic site and then 'bounce' out), extended navigation (people spend as much time finding their way around as actually viewing results), horizontal "power browsing" through titles, contents pages and abstracts, squirreling behavior[10], and little time spent in evaluating information. At the same time, it has been observed that users find

[8] An expectation that certain kinds of resources and information would be found in library/academic sources and not in search engines

[9] Hildreth (1995) explained the popularity of the Web by its exploratory "browser" interfaces that support many users' preference of "action and encounter" to "reflection and analysis."

[10] Saving information in form of downloads for later use, particularly free content (though it is rarely re-visited by the downloader).

system functions supporting user tasks involved in resource discovery by subject — subject clouds, keyword search and subject search, collocation by subject options, content summaries/abstracts — helpful in searching (Zhang & Salaba, 2007b).

Such patterns of user searching as the use of Boolean operators and controlled vocabulary in online catalogs have been analyzed in a number of studies. Boolean search was found to be ineffective, not only because the majority of library users — even highly educated ones — experienced difficulties with Boolean logic concepts, but also because the execution order of Boolean commands was not standardized across different OPACs (e.g., Borgman, 1996). Moreover, Allen (1991) found that performance is improved in systems that do not require using Boolean operators for complex queries. A number of studies have shown that although both natural language/keyword search and controlled-vocabulary search produce effective retrieval results, users of OPAC tend to search more often by keyword than by any other type of search (e.g., Fidel, 1988, 1992; Curl, 1995; Hildreth, 1997; Muddamalle, 1998).

Polyrepresentation of information objects (Ingwersen, 1994) where the system contains multiple sets of metadata (e.g., both authoritative metadata containing the relatively stable and defined attributes, and user-generated context metadata such as tags) has been viewed as a possibility for improving subject access to large databases. The library community is starting to work with user tagging in order to expose library resources via new routes and to allow users to interact with resources in new ways. For example, the Library of Congress is experimenting with Flickr to get help from Flickr users to tag and describe part of its extensive collection of photographs (e.g., Raymond, 2008). A number of libraries and other cultural institutions have joined the Flickr Commons or created non-Commons Flickr photostreams. Moreover, a recently released report by the Library of Congress on the future of bibliographic control recommends to

the “library community as a whole and its close collaborators” an integration of the user-contributed data (tags) into library catalogs (*On the record,* 2008, p. 32).

1.4.3 The Information-Seeking Behavior of Scholarly Historians

A number of investigations into the information seeking behavior of historians and their interaction with information retrieval systems have been conducted. The findings are summarized in this section.

Searching — an integral part of any research project — occurs at various stages of historians’ research work. The purposes of searching include: obtaining support for an argument, finding out about something that is unclear, seeing if there are any new developments, following up on leads, and finding a work to incorporate, etc. It has been found that historians in general search mostly for primary — most often unpublished — sources. While relying on books much more than on journals (e.g., Stone, 1982), they frequently use other text-based objects (e.g., diaries, wills, letters, manuscripts), visual representations (e.g., photographs, portraits, architectural drawings, films), and even three-dimensional objects such as toys (Case, 1991). Like other humanities scholars, historians rely heavily on their own personal collections while using a variety of material rather than a well-defined core of material (Reynolds, 1995). Historians also search in other domains including: philosophy, anthropology, art history, criticism, literature, statistics, sociology, criminology, geography, and physiology (Case, 1991).

Historians often use electronic means to locate primary materials for their research; visiting Web sites of known repositories is a more frequent behavior than using search engines (Tibbo, 2003). Historical researchers greatly value digitized archival databases, which often help them locate materials that they have sought for years (Duff & Johnson, 2002). The use of digital

libraries peaks during the initial stages of a research or teaching project (Buchanan et al., 2005). A study conducted by Garrett (2007) revealed the importance of controlled-vocabulary subject headings (LCSH) for access to historical materials in digital libraries and demonstrated the value added by subject headings in a full-text environment.

Duff and Johnson (2002) observed the following search techniques used by historians in exploration of archival information:

1. collecting names of individuals and organizations related to research to use them later as pointers to specific collections and archives
2. using keywords found and written down when studying finding aids at the initial stage of archival research
3. "provenance search method," when information needs are connected to functions and activities of an organization studied.

They also found that contextual information (e.g. knowledge of relationships among the documents or the way archival records are organized) is critical for the search: "The totality of the records provides information that no individual record can. Historians must comprehend the records in their context rather than as separate disembodied items. Without this context information, the historian could easily misinterpret the meaning or significance of the information in an individual record" (Duff & Johnson, 2002, p.487).

According to Bates' search term taxonomy (1996), key query term types that appear in the searches of humanities scholars, for example, historians, include names of individuals, geographical names, chronological terms and discipline terms (concepts within the field). When searching electronic catalogs (Buchanan et al., 2005), humanities researchers use a variety of search types, with known author-title search being the least problematic. Success in more

uncertain types of searches — e.g., a conceptual /discipline term search — heavily depends on the level of searcher's domain knowledge and experience in using a particular digital library. Subject classifications were almost never used by academic searchers in Buchanan's study because the scholar's conceptual models usually differed from that represented in the classification scheme.

Browsing has traditionally been a crucial way of finding information. Recent research provides evidence that humanities scholars, including historians, still heavily rely on browsing (e.g., Ellis & Oldman, 2005). However, although often engaged in browsing in physical libraries and recent journal issues, humanists and social scientists in the Buchanan et al. (2005) study preferred searching to browsing in a digital library environment due to the lack of call number browsing capabilities.

1.5 Digital Collections, Free-Text and Controlled-Vocabulary Collection Metadata, and Users

1.5.1 Digital Collections

As noted in the introduction, digital collections rather than individual items are the focus of this dissertation research. This study seeks to fill the existing gap in research, which had focused on item-level user interaction with metadata, and also to provide a better understanding of the role of collection-level metadata and of the user perspective on information discovery in aggregations of digital collections.

Characterizations of digital collections vary widely in the literature. As summarized by Cleveland (1998), digital collections:

- include materials that exist both within and outside the physical and administrative bounds of any one digital library,
- ideally provide a coherent view of all of the information contained within a library, no matter its form or format,
- serve particular communities or constituencies, although those communities may be widely dispersed throughout the network.

The CIDOC object-oriented conceptual reference model (International Council of Museums/CIDOC, 2007) emphasizes the purposive nature of digital collections defined as "aggregations of physical items that are assembled and maintained … by one or more instances of Actor over time for a specific purpose and audience, and according to a particular collection development plan. Items may be added or removed from a Collection in pursuit of this plan." The simple definition of a collection as "any aggregation of individual items (objects, resources)" (Johnston & Robinson, 2002) does not contain limitations as to the form and nature of items in a digital collection — either digital items as surrogates of physical items or "born-digital" content objects. This definition views catalogs as tantamount to a collection, and remains neutral to the collection size, which can be as small as one item.

Lee (2000) proposed expanding the concept of collection to represent a group of documents, regardless of format, medium, and ownership. She also points that collections coexist in multiple layers (i.e., some collections are subcollections of others). Cole and Shreeves (2004) identified several criteria for operationalizing the definition of a digital collection: thematic cohesiveness (e.g., by topic area, holding institution, type of materials), searchability as a distinct collection, and a unique point of entry (URL).

It has been noted that there never was a universally accepted definition of the term "collection" in the LIS field (e.g., Lee, 2000; Hill et al., 1999). In a digital environment, where collection practices are changing, and the principles that guide collection building are evolving, a common understanding of the concept seems more elusive than ever. The fluidity of digital content seems to have resulted in the concept of the collection becoming more abstract than in the past (Manoff, 2000), even though the dynamic nature of collections has always been accepted as part of management, as repositories of all kinds commonly add and delete objects (Currall, Moss, & Stuart, 2004). Researchers (e.g., Johnston & Robinson, 2002) emphasize the transient nature of digital collections and the fact that items in such collections are often dispersed across multiple physical locations. Collections have been conceptualized with a certain degree of polarity: as contexts for information seeking (Lee, 2000) or as bodies of raw materials for interpretation and presentation by user (Lynch, 2002). Traditional user-based collection criteria are still being considered in building digital collections (Lagoze & Fielding, 1998). Some researchers emphasize permanence of digital collections, while others stress the transitory nature of such collections.

An investigation into how developers conceptualize their own digital collections (Palmer, Knutson, Twidale, & Zavalina, 2006) revealed high variability and ambiguity in the collection construct. Most resource developers consider their digital resource as multiple collections, while others think their digital resource could be considered either one or several collections. Many do not have a firm idea of how many collections they are creating. When a resource is conceptualized as one collection, often it is due to an archival perspective on an integrated whole. Archivists' notions of "artificial" and "organic" collections retain relevance beyond the archival community, and the collection genre of "exhibit" is being adopted outside the museum

community. Purpose-based terms such as “displays”, “tours”, “tools”, “lessons,” and others are also used to describe digital resources. Digital resource developers’ use of the term “collection” is surprisingly diverse, and some explicitly remark that the term is fuzzy or problematic.

Some researchers have stated that collections might play an important role in information seeking (Allen & Sutton, 1993). However, the lack of research on collection structures (i.e., components and the organization among the components) and their effect on information seeking and use impedes effective system and service design and results in the lack of user-centered approaches in structuring collections (Lee, 2003). Lee (2000) has suggested that collections and collection structures may facilitate or hinder information seeking, while information professionals’ and users’ criteria for conceptualizing and structuring collections differ. Her studies (2003, 2005) demonstrate the usefulness of collections and subcollections (with certain subcollections not defined by the library as distinct structures) in information seeking of academics. The valuable functions provided by collection structures include collocating sources, selectivity, narrowing the search scope to increase precision and ease of use, presenting choices and assisting in information need clarification. Such functionality becomes especially important in a federated digital resource environment. The role that collection structures, particularly collection-level metadata, play in scholarly user interaction with aggregations of digital collections constitutes a central part of my dissertation research.

1.5.2 Collection Metadata

Collection metadata has a vital role to play in facilitating access, especially in the digital environment. Collection-level descriptions, defined as “a structured, open, standardized and machine-readable form of metadata providing a high-level description of an aggregation of individual items” (Macgregor, 2003), provide an added level of descriptive granularity:

important relational (Macgregor, 2003) and contextual information (Miller, 2000), functional both for the user and the institution. Contextual metadata has long been recognized in the archival community as being central to facilitating access to documents in archival collections (e. g., Bearman, 1992). Best practice recommendations for Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) data provider implementations and shareable metadata stress the importance of retaining context when aggregating item-level metadata and the necessity of expressing and sharing descriptions of the collections to which the resources described by item-level metadata belong (e.g., Shreeves, Riley, & Milewicz, 2006). Research using metadata harvested from the Committee for Institutional Cooperation (CIC) demonstrates that linking item-level and collection-level metadata can help produce higher retrieval rates for item-level descriptions, re-contextualize orphaned items by including key access points that might be lacking in item-level metadata into collection-level metadata, and facilitate browsing behavior familiar to humanities scholars by providing links from item-level records to the relevant collection-level records (Foulonneau et al., 2005).

Geisler et al. (2002) claim that relational attributes specifying relations between a given collection and its various sub- and super-collections will be essential in collection-level descriptions, for discovering resources within single repositories, across institutions, and across different domains. These attributes have a capacity to "greatly improve the navigability of the [digital library]" (p.217). Similarly, Heaney (2000) distinguishes between a unitary collection description, which "consists only of information about the collection as a whole and does not provide information about the individual items within it" and analytic collection description, which "consists of information about the individual items within [a collection] and their content" (p.18).

At the item level, Dublin Core is one of the two most widely used metadata schemes in digital libraries, surpassing even the MARC standard traditionally used in online catalogs (e.g., Palmer, Zavalina, & Mustafoff, 2007). The Dublin Core Collection Application Profile (DCCAP) has been developed to guide collection-level metadata creation. Like the item-level metadata elements, collection-level metadata elements can be subdivided into two kinds based on how they are encoded — formalized metadata usually expressed through controlled-vocabulary terms (e.g., subject, format, object type, etc.) and free-text metadata (e.g., title, abstract, notes, etc.).

The *Description* field, defined by DCCAP as a required "free text summary description of the collection"[11], has been an integral part of collection-level metadata, providing important human-readable contextual information for users. The DCCAP does not prescribe what should be included in the collection-level free-text *Description* field, however subjects of a collection are suggested as possible content: "Although a description might contain detailed subject-specific information, at least part of the description should be understandable by an end-user with no specialist knowledge of the subject area." The Dublin Core Metadata Elements Set for item-level metadata[12] provides a slightly more detailed definition and some guidelines as to the contents of the mandatory *Description* field: "An account of the content of the resource", "may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content." The Dublin Core Usage Guide[13] recommends limiting the length of *Description* field to "a few brief sentences."

[11] http://dublincore.org/groups/collections/collection-application-profile/#coldctermsabstract.
[12] http://dublincore.org/documents/dces.
[13] http://dublincore.org/documents/2001/04/12/usageguide/sectb.shtml#description.

The usage guides created by different communities for their own needs suggest that collection- and/or item-level *Description* information should "be helpful to users attempting to discern the usefulness of a resource to their research needs" (*NCSU Libraries Core 1.0 Metadata Element Set Best Practices*, 2007), and provide information that is not covered by other metadata elements or "supplement, qualify, or explain" information in other metadata elements (*Cataloging Cultural Objects*, 2008). Usage guides have recommended providing information about:

- "salient characteristics and historical significance of the subject, function, and significance of the work," work's "relationship to other works, its style, and any aspects of it that might be either disputed or uncertain" (*Cataloging Cultural Objects*, 2008),
- types of materials included in collection, associated dates, "names, dates, and biographical identification of persons and names of corporate bodies significant (by quality and/or quantity of material) to the collection," specific phases of career/activity of the major person/body responsible, geographical areas, events, topics, and historical periods with which the materials in the collection deal, and "particular items of extraordinary interest,"[14]
- "provenance, distinguishing features, inscriptions, the nature of the language of the resource, and/or history of the work" (*OSU Knowledge Bank Metadata Application Profile*, 2006).

The broader cataloging/metadata community has developed detailed guidelines for creating descriptive summary notes in MARC-format item-level records, which might be useful

[14] Webform for creating collection records in *National Union Catalog of Manuscript Collections* http://www.loc.gov/coll/nucmc/lcforms.html (paragraph #10).

in thinking about encoding of the collection-level *Description* field content as well. The guidelines created by *Online Audiovisual Catalogers'* Cataloging Policy Committee (2002) recommend including such elements as "unique features" or "distinguishing features", "user interaction", "specific effects" (e.g., laser display or animation), and "history of the work," when describing individual items. These guidelines also mention including audience information when creating summary notes in item-level records for motion pictures and video recordings. For describing archival materials — normally represented as collections — OLAC guidelines recommend inclusion of summary note information about "specific types and forms of materials present", "reason and function of the collection", "significant people, places, events and topics covered," "span of dates covered by collection", "typical and unique characteristics of the collection," and "consequences, products, and results of the events documented."

Both free-text and structured formalized collection-level metadata is being created to facilitate resource discovery in aggregations of digital collections. Due to their cultural heritage focus and strong representation of primary sources of information, many such aggregations are of special interest to scholarly historians. However, little research into digital information seeking behavior of this user group has been conducted to date. The next subsection summarizes existing research.

1.5.3 Historians as Users of Digital Collections and Aggregations

Scholars in general are one of the major audiences for various aggregations of digital collections. For example, 54% of the survey respondents of IMLS-funded digitization projects (*Assessment of End-User Needs…*, 2003) name scholars as a target audience. Academic historians and history enthusiasts have been a target user group for a number of aggregations of digital content: American Memory, American Social History Online, and Opening History, to

name a few. Although the amount of digital content of interest to historians is increasing, there has been little research done concerning the use of such resources by these target audiences.

The usability studies conducted by the developers of American Social History Online aggregation of digital collections, conducted with history faculty members and doctoral students (Harum, 2008), demonstrate that historians:

- are interested in image items more than text-based items, thus, heavily relying on thumbnails for browsing and selection both at the collection- and item- level,
- actively use the browsing feature (particularly, timeline, chronological browse, and interactive map browse) for teaching and research,
- find list-like subject browse feature "too much to look at,"
- need the personal and geographic name browse capability,
- value the use of facets (e.g., state, city, genre, decade, collection, language, media type, etc.) to limit search results set,
- want to be able to go from item search results to relevant collections,
- do not tag digital content and do not trust other users' tags.

Another recent study (Wu & Chen, 2007) collected somewhat similar results. The authors found that, while interacting with full-text digital collections, history graduate students:

- want personal and geographic name search capabilities, and search limit by date,
- suggest including biographical dictionaries, gazetteers, or authority files for personal and geographic names,
- are interested in hyperlinks to related documents

- value metasearch capability, more personalized settings and Web 2.0 capabilities such as a space for discussion and user reviews on digital items,
- use search feature more than browse feature, while preferring basic search to complex.

Many of the functions which historians find useful in interaction with digital collections and aggregations have been implemented by Opening History, an aggregation of cultural heritage digital collections developed for historians as the primary user group. Section 1.6 introduces the Opening History aggregation, which was selected as the major target for my dissertation research, and the pilot studies conducted on its predecessor's − IMLS DCC Collection Registry − collection-level metadata application and use, which both informed the research questions and research design of this study.

1.6 Opening History Aggregation of Digital Collections

The Opening History aggregation is one of several large-scale U.S.-based federal-level aggregations of digital content created in recent years. The Digital Collections and Content (DCC) project, funded by the Institute of Museum and Library Services (IMLS), has been in operation at the University of Illinois at Urbana-Champaign since 2002. After developing a collection description metadata schema, the DCC project created an IMLS DCC Collection Registry of digital collections funded through the IMLS National Leadership Grant (NLG) and built by cultural heritage institutions since 1998. Selected collections funded through the Library Services and Technology Act (LSTA) grant have been included since 2006. The Opening History aggregation, which includes digital collections focusing on United States history, regardless of funding sources, was started as part of the DCC project in Fall 2008, with 227

collections, and has been rapidly growing since then. As of July, 2010, the Opening History aggregation consists of 864 digital collections. Digital content from approximately 20% of digital collections has been harvested into an item-level repository adding over a million item-level records. The types of digital content in the Opening History include image, text, physical object, sound file, interactive resource, moving image, and dataset. Military history, Native American history, and transportation history are some of the major subject strengths of the Opening History aggregation.

The collection metadata scheme used in the Opening History aggregation, inherited from the IMLS DCC Collection Registry, was developed based on the UKOLN Research Support Libraries Programme (RSLP) scheme[15] and later aligned with the Dublin Core Collection Application Profile.[16] The DCC Collection metadata scheme (Figure 3) describes four entities: the digital collection itself, the grant project responsible for collection, the institution responsible for collection, and the person(s) responsible for the administration of the digital collection. For describing a collection per se, the scheme provides nineteen general attributes (name of the collection, alternative title, objects represented, collection URL, creator, interactions with digital collection, format of digital items, language, size of collection, frequency of additions, supplementary materials, audience, access restrictions, rights, collection development policy, alternative access, notes, custodial history, and date items created), four topical attributes (topic, [free-text] description, geographic coverage, and temporal coverage), four attributes describing relationships with other collections (parent collection, sub-collection, source physical collection, and other associated collection), and four attributes describing relationships with digitization

[15] http://www.ukoln.ac.uk/metadata/rslp
[16] http://dublincore.org/groups/collections

projects, institutions, and administrators (grant project, hosting institution, contributing institution, and administrator).[17]

Mainly to support general subject browsing, topics of the digital collections in the Opening History aggregation are indexed with the Gateway to Educational Materials (GEM)[18] a subject vocabulary considered suitable for browsing databases in a cultural heritage domain. While the *GEM Subjects* is a required collection metadata field in DCC collection metadata scheme, three more optional metadata fields are intended in DCC collection-level metadata schema for subject indexing: *Subjects* (for terms from controlled vocabularies other than GEM or uncontrolled keywords), *Time Period*, and *Geographic Coverage.*

Figure 4 displays three search options provided in the Opening History aggregation: *simple* (searches both items and collections), *advanced* (searches only items by keyword/phrase anywhere, by author/artist's last name, and by title/subject word(s), with a possibility to limit results to selected collections), *"search collections only"* (simple keyword search in all fields in collection metadata records, with a possibility to limit search to types of objects in collections: dataset, interactive resource, physical object, text, image, moving image, sound, and unknown). In addition, users can browse collections in the Opening History aggregation by subject, object type, place, collection title, hosting institution, and, for LSTA- or NLG-funded grant project. The item-level browse functionality is not currently available.

[17] General overview of the IMLS DCC collection description scheme is available at: http://imlsdcc.grainger.uiuc.edu/CDschema_overview.asp while detailed description of the scheme's elements can be found at: http://imlsdcc.grainger.uiuc.edu/CDschema_elements.asp

[18] http://www.thegateway.org/about/documentation/gem-controlled-vocabularies/vocabulary-subject

1.6.1 Pilot Studies

In the first pilot study (Zavalina, 2007), transaction log data from the IMLS DCC Collection Registry for the period of February-September 2005 was analyzed qualitatively and quantitatively to find answers to the following research questions:

- What are the typical collection-level user search categories? How adequate is FRBR entity-relationship model for categorization of user search terms?
- What is the distribution of the two major collection-level search types (subject and known-item)?
- What are the quantitative characteristics/patterns of user search queries?

Content analysis of 936 user keyword search queries extracted from transaction logs of collection level searches was conducted to determine search category and type. Findings showed heavy use of three FRBR Group 3 search entities: *object* (24%), *concept* (21%), *place* (15%), and one FRBR Group 2 entity: *individual person* (13%). A surprisingly low level of *event* (FRBR Group 3 entity) searching (4%) was found. The study also discovered two additional search categories that are not currently covered by the FRBR model: *ethnic group* and *class* of *persons.* With respect to search types, broadly-defined subject searching (including *concept, object, place, event, ethnic group,* and *class of persons* categories) was prevalent, at 75%, which is unusually high for catalog use / transaction log analysis studies. Findings with respect to the search query characteristics demonstrated high variability of user keyword queries in complexity and length. The number of words in a query ranged from 1 to 7, with the vast majority consisting of one or two words, the average query length for the whole sample constituted 1.67 words per query. A rather low overall average frequency of term use (1.4) was observed. These findings helped to refine the FRBR-based list of the search categories to be used in dissertation research

by inclusion of *ethnic group* and *class of persons* categories. The pilot study revealed that high proportion of the searches cannot be adequately categorized when assigned to a single search category; this finding served as a basis for decision to allow assigning user search queries to multiple categories in the main study. Results also called for further research into the possible qualitative (e.g., distribution of search categories) and quantitative (e.g., query length and frequency) differences between item-level and collection-level search in aggregations.

The second pilot study (Zavalina et al., 2008a, 2008b) focused on a detailed exploratory content analysis of collection-level metadata in 202 digital collection records then in the IMLS DCC Collection Registry. The free-text *Description* field was found to add essential subject information to a record by providing more specific coverage than controlled-vocabulary fields intended for subject indexing (*GEM Subjects,* [alternative] *Subjects, Geographic Coverage*, and *Time Period*). The findings confirm that an important role is played by the free-text *Description* field in providing information about collection subjects (91%) often not provided elsewhere in the records (67%), including spatial (60%) and temporal (50%) coverage of collections. A consistent indication of types of object in a collection (75%), sometimes not provided elsewhere in the records (19%) was also observed in the content analysis. Statistical analysis also demonstrated that the length of a *Description* field appears to have an effect on occurrence of indications of the subjects beyond those covered by specialized fields. These findings demonstrated that free-text *Description* metadata field is a rich source of subject-specific (topical, geographic, temporal, genre) information about a digital collection and suggested that the richness of this field (e. g., number of various collection characteristics encoded in it) could be an important component of overall collection metadata richness. Collection metadata richness was used in this dissertation research as the measure of the value of collection metadata in

subject access to digital collections in aggregations. These findings also posed a question if and how the subject-specific information encoded in controlled-vocabulary fields complements similar information recorded in free-text *Description* field.

The third pilot study conducted in May 2008 analyzed the result sets for 100 collection-level user searches (a cluster sample derived from IMLS DCC Collection Registry's transaction logs) replicated in IMLS DCC Collection Registry, with the goal to determine which collection metadata fields provided the matches to the user search terms. This pilot study tested the adaptation of the Gross and Taylor's (2005) item-level metadata analysis technique in the context of collection-level metadata. The analysis determined that 45% of searches return between 1 and 96 search results with matches in various collection metadata fields. As shown by Table 1, free-text collection metadata, such as *Description, Notes, Title, Copyright and IP Rights*, etc., was found to play an important role in collection-level information discovery. For example, in 58% of successful collection searches, the search query could be satisfied only through the free-text *Description* field. At the same time, the information contained in mostly controlled-vocabulary collection metadata fields intended for subject representation, such as *Subjects, Objects, Geographic Coverage,* and *Time Period,* provided a match to a significant proportion of user search terms. Results of this pilot study suggested that applying a variety of free-text and controlled-vocabulary metadata fields with mutually complementary values can be important in satisfying the user searches. This finding was instrumental in refining the definition of collection metadata richness used in this dissertation research.

1.7 Figures and Tables

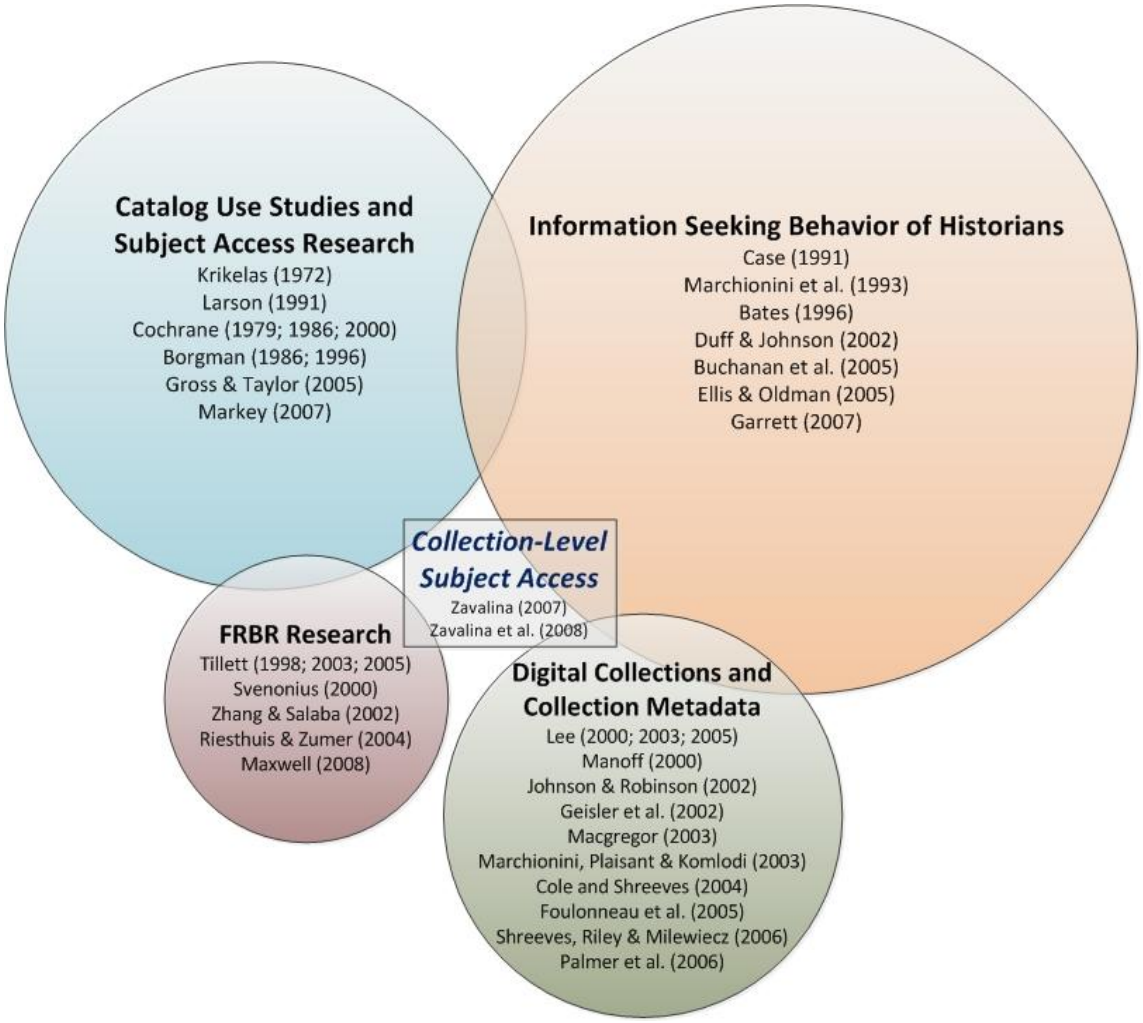


Figure 1. Position of collection-level subject access research among related areas of research

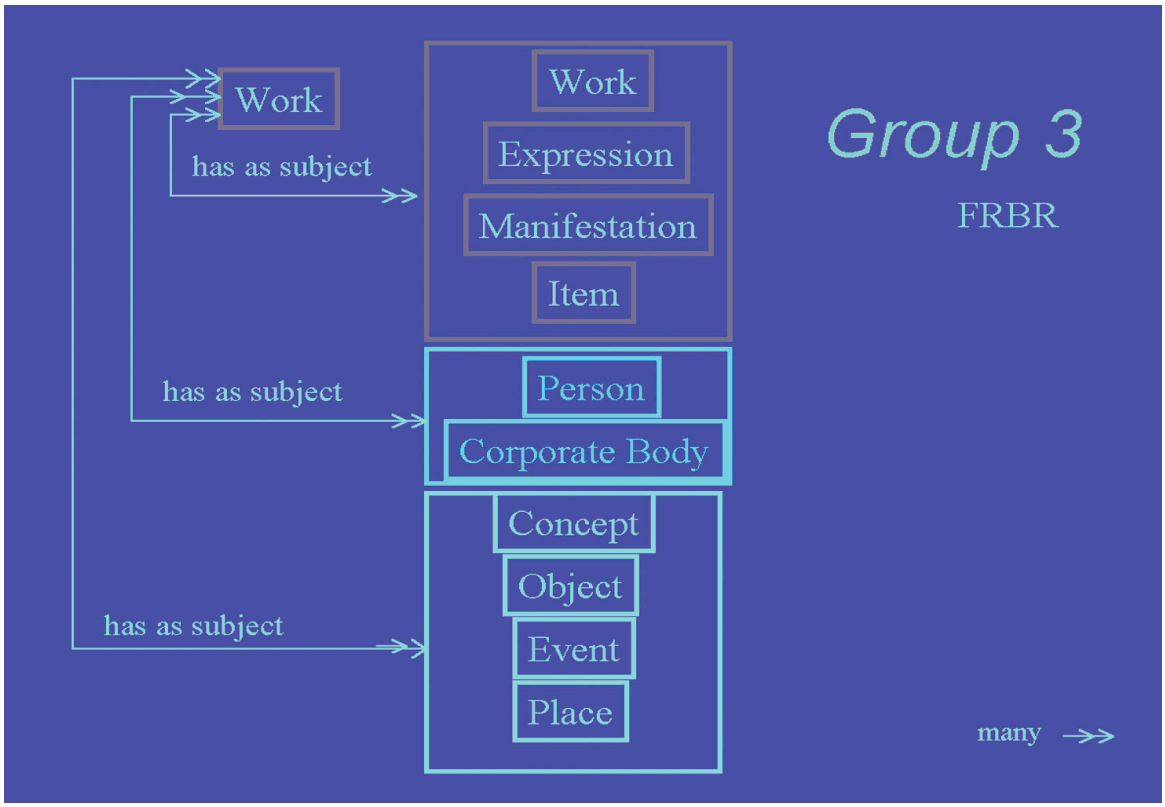


Figure 2. Subject entities and relationships in the FRBR model
[From Tillett, B. (2004). What is FRBR? http://www.loc.gov/cds/downloads/FRBR.PDF, p.3].

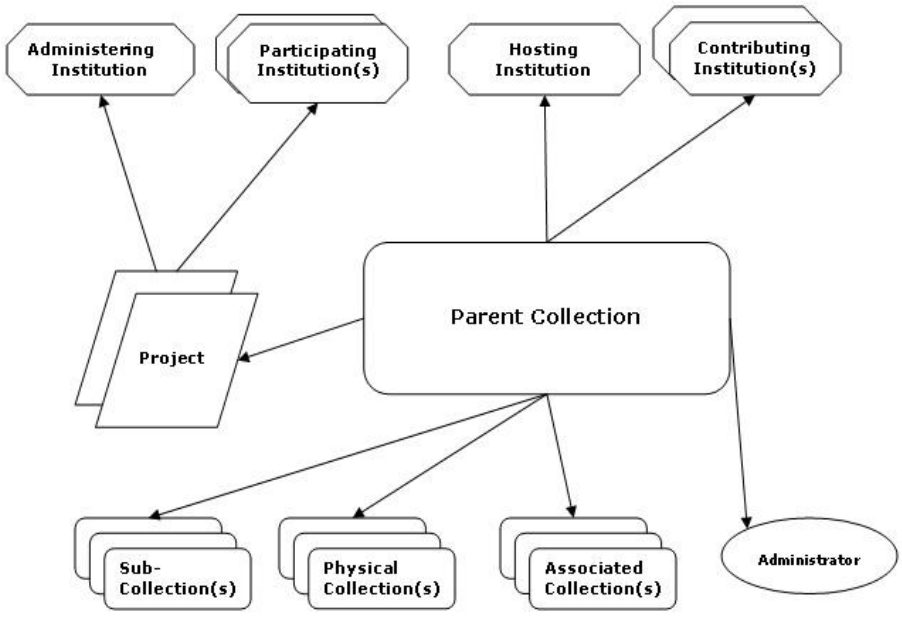


Figure 3. DCC collection metadata scheme

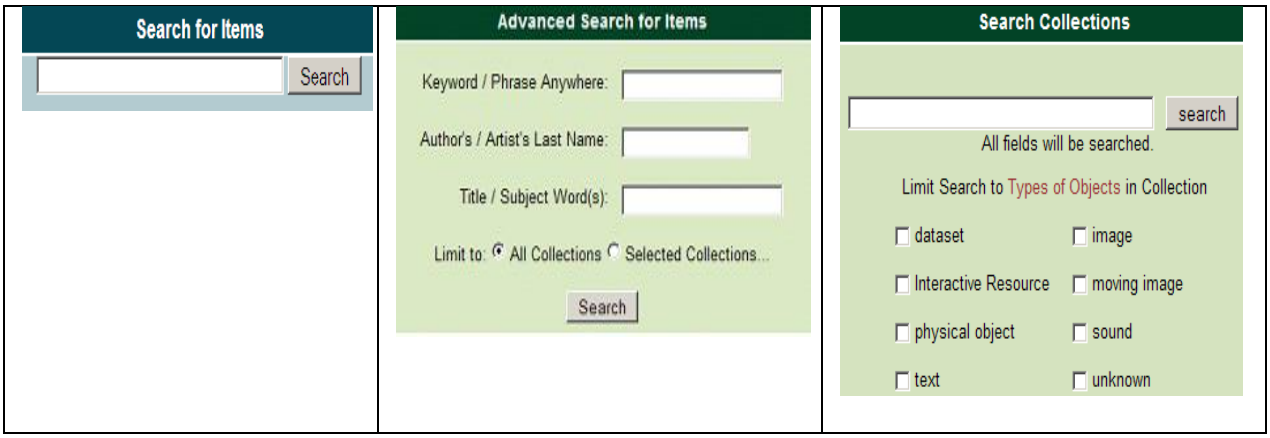


Figure 4. Search options in Opening History aggregation

Collection metadata fields containing match(es) to user search queries	Nature of the field	% of successful searches
Description (overall) ONLY *Description*	Free-text	80 58
Subjects[19] (overall) ONLY *Subjects*	Controlled-vocabulary	58 36
Title	Free-text	44
URL	Free-text	22
Notes (overall) ONLY *Notes*	Free-text	20 16
Objects	Controlled-vocabulary	13
Copyright & IP Rights	Free-text	13
Contributing Institution	Free-text	11
Geographic Coverage	Controlled-vocabulary	9
Size	Controlled-vocabulary	9
Hosting Institution	Free-text	7
Audience	Free-text	4
Alternative Access	Controlled-vocabulary	2
Format	Controlled-vocabulary	2
Temporal Coverage	Controlled-vocabulary	2

Table 1. Collection metadata fields with matches to user search terms (pilot study of IMLS DCC Collection Registry)

[19] Including *GEM Subjects* and alternative *Subjects* fields.

Chapter 2. Problem Statement and Research Questions

2.1 Introduction

Over the last fifty years, researchers in library and information science have become increasingly interested in examining the intellectual access to information provided to users by existing information organization systems: catalogs and bibliographic databases of various kinds (e.g., Jackson, 1958; Lipetz, 1970; Tagliacozzo & Kochen, 1970). Research methods employed in these studies have included interviews and observations, panel discussions, surveys, and transaction log analyses. This research, accelerated by the development of online public access catalogs, revealed that subject, along with the "known-item," was a primary access point for users searching for information in bibliographic databases, but that users experienced problems with this type of approach (Krikelas, 1972; Matthews, Lawrence, & Ferguson, 1983; Borgman, 1986, 1996; Larson, 1991a). As discussed above, factors contributing to the effectiveness of subject access include quality and application of controlled vocabularies (e.g., Cochrane, 1986, 2000), user subject domain knowledge / conceptual knowledge (e.g., Bates, 1977; Allen, 1991; Borgman, 1996; Markey, 2007), user understanding on how to utilize the information organization system — semantic knowledge and technical skills (Borgman, 1996), and procedural knowledge (sources to search and the most efficient order for searching) (Markey, 2007).

Research to date has focused on the item-level information discovery and access. In the last decade, however, the creation of federated/aggregated collections and collection registries on the Web has connected distributed digital content and provided users with the ability to search across collections. While some of these aggregations in United States and abroad (e.g., OAIster,

Europeana, etc.) do not use collection-level descriptions, others consider collection metadata important for providing context for the digital items harvested from distributed collections. In the United States, American Memory, Opening History (and its predecessor IMLS Digital Collections and Content Collection Registry), and National Science Digital Library are the largest aggregation projects among the hundreds existing[20] and have accumulated significant experience in describing digital collections. Overseas, The European Library is one of the most prominent large-scale international aggregations. The collection-level descriptions provided by these aggregations, in general, and subject representation in particular, need to be analyzed and compared to provide a better understanding of collection-level metadata application patterns (and indication of best practices).

User interactions with large-scale aggregations, which are often comprised of distributed resources and organized by collections, are likely to differ from user interactions with library catalogs, and even from standard web searching. However, while a number of researchers have analyzed the digital resource developers' perspective on defining and describing digital collections to facilitate resource discovery (e.g., Hill et al., 1999; Lee, 2000; Currall, Moss, & Stuart, 2004; Palmer et al., 2006), little attempt has been made to examine the user perspective on information search and discovery in aggregations of digital content (e.g., Twidale & Urban, 2005). For example, no research has yet focused on the role subject representation plays in user interaction with aggregations of digital collections.

Additionally, the FRBR entity-relationship model of the bibliographic universe (IFLA, 1997, 2008) has been increasingly influential in thinking about the ways to organize descriptive

[20] See for example this extensive list of digital content aggregations http://oedb.org/library/features/250-plus-killer-digital-libraries-and-archives.

metadata in databases. However, the model needs to be verified and validated against real data in different communities (Zhang & Salaba, 2007a), while a specific need exists for researching how subject searching is done in practice (Riesthuis & Žumer, 2004). In particular, the understanding of how users search in the aggregated digital content environment would help update and align the model. Discovering how the FRBR model fits collection-level subject searching by scholarly historians would contribute to such understanding.

2.2 Research Questions

This dissertation research is aimed at bridging the gap identified above by addressing the overarching research question: ***How does collection-level metadata mediate scholarly subject access to aggregated digital collections?*** Consequentially, how rich does the collection-level record need to be in order to be useful for scholarly subject access? How crucial is manually-created formalized collection-level metadata for scholarly subject access? Can free-text metadata alone satisfy most of the collection-level user requests?

The following specific questions — components of the overarching research question — have been pursued in the course of this study:

- What is the variation in richness[21] of collection-level subject metadata across collections in aggregations of digital collections?

[21] Details on measuring collection metadata richness are provided in the section 3.2 of this thesis.

- How do scholarly users of cultural heritage aggregations approach collection-level information discovery?
- Which collection-level metadata fields provide scholarly users with the most valuable information to meet their needs?
- How does collection-level user search data fit the FRBR model of bibliographic universe?

The research questions are based on the underlying assumption that scholars — and scholarly historians in particular — constitute a major audience of the aggregations of cultural heritage digital collections. In addition, the following conjectures regarding the role of collection-level metadata in facilitating scholarly collection-level subject access and resource discovery in aggregations of digital collections, based on the literature reviewed in the Chapter 1 of this thesis, have informed the research questions:

- collection-level metadata retains and provides important contextual information when aggregating item-level metadata (Shreeves, Riley, & Milewicz, 2006; Miller, 2000),
- collection-level metadata helps narrow the search scope to increase precision and ease of use (Lee, 2003, 2005),
- collection-level metadata presents choices and assists in clarification of information need (Lee, 2003, 2005),
- collection-level metadata provides important relational information specifying relations between a given collection and its various sub- and super-collections (Macgregor, 2003; Geisler et al., 2002),

- collection-level metadata facilitates the browsing behavior which is familiar to humanities scholars by providing links from item-level records to the relevant collection-level records (Foulonneau et al., 2005).

The conjectures listed above, which, with the exception of Lee's studies are not based on the analysis of user interactions with aggregations of digital collections, has served as a context and a point of comparison with the user interaction data collected in this dissertation study.

Although virtually no research exists to date regarding patterns of user behavior in collection-level aggregations of digital collections, the following conjectures, based on observations from the pilot studies, have also informed the user interaction part of this dissertation study:

- Sholarly users of aggregated digital collections might search at the collection level differently than at the item level (e.g., formulate relatively broad search terms, give preference to certain categories of search terms such as concepts or objects, choose browsing over search, etc).
- Confusion may exist in making distinctions between searching for individual items in a collection and searching for collections in an aggregation. Such ambiguity can cause unjustified preciseness and narrowness in collection-level search terms.

Chapter 3. Method

3.1 Research Design Overview

Three large-scale aggregations of digital collections were identified to investigate collection-level subject representation and subject access. The Opening History aggregation has been the major target of this study. Two other aggregations have been used for comparative purposes, one U.S.-based − American Memory − and one international, − The European Library.

To address the research questions, the research process has been divided into two phases:

- PHASE 1. Collection metadata analysis: analysis of current approaches to collection-level subject representation in the three aggregations.
- PHASE 2. User interaction analysis: analysis of scholarly users' perspectives on collection-level subject access.

Table 2 outlines the methods of data collection and analysis employed in each of the research phases. The research methods are further detailed in sections 3.2-3.3 of this thesis.

3.2 Phase 1. Collection Metadata Analysis

The first phase of data collection and analysis focused on the collection-level metadata records provided by the three aggregations. Phase 1 sought to answer the first of the two overarching research questions:

What is the variation in richness of collection-level subject metadata across aggregations of digital collections?

Although no consensus on defining the richness of metadata (either item-level or collection-level) exists,[22] for the purposes of this study the richness of collection-level metadata was defined as the combination of three measures:

- variety of collection-level metadata elements used in collection-level description.
- the length of the values encoded in subject-specific metadata elements (e.g., the number of words in the free-text *Description*, *GEM Subjects*, *Subject*, *Time Period* and *Geographic Coverage* field).
- the number of collection properties (e.g., uniqueness, provenance, subject, object, navigation and functionality, etc.)[23] represented in the free-text collection-level metadata (e.g., *Description* field).

Content analysis has been used as the major method at this stage. Content analysis is a widely used method of study in the social sciences and LIS (reviewed by Allen & Reser, 1990; Weare & Lin, 2000). The two main types of content analysis are quantitative and qualitative. In both types, the analysis usually begins with the manual exploration of a text or a dataset to identify language patterns that correspond to the processes and concepts under investigation, which are often referred to as categories. The procedure for finding and labeling categories is usually referred to as coding. The next step of content analysis involves selection of the unit of analysis appropriately representing categories studied. A unit of analysis can range in size from a single word to the whole text. Although the unit of analysis is usually predefined, some researchers (e.g., Henri, 1992) emphasize the need for thematic units or meaning units, the size of which may vary to represent studied phenomena more accurately.

[22] According to Duval et al. (2002), the richness of metadata descriptions should be "determined by policies and best practices designated by the agency creating the metadata, and those policies and practices will be guided by the functional requirements of services or applications."
[23] See pilot study no. 3 (section 1.6.1 of this thesis).

Categories are used differently by the two approaches. In quantitative content analysis, after coding all instances of studied categories, researchers usually count the number of instances in each category and apply different statistical tests to determine the dominance of a particular category, to identify relationships between categories, and to compare results across different datasets. In qualitative content analysis, after the coding procedure is completed, researchers do not run statistical tests but rather try to deduce trends or specific phenomena from the coded text.

For this dissertation research, a combination of qualitative and quantitative content analysis was used for analyzing collection-level metadata in three large-scale aggregations of cultural heritage digital collections. The units of analysis range depending on the specific research question being answered: from a phrase or sentence to entire contents of a field in collection-level record, to a whole collection-level record.

The entire population of collection-level records in Opening History aggregation http://imlsdcc.grainger.uiuc.edu/history, which included 496 collections as of February 1, 2009, was analyzed. Eight collections with *Description* field values that duplicate values of *Description* fields in other collections were excluded from the sample. Hence, the Opening History sample size was reduced to 488.

For comparative content analysis of collection records in two other aggregations — American Memory[24] and the European Library[25], the following sampling procedure was applied:

- Collection-level metadata for all the digital collections that are also part of the Opening History aggregation was selected from the other U.S.-based aggregation —

[24] http://memory.loc.gov; included 138 collections as of October 28, 2008
[25] http://www.theeuropeanlibrary.org; included 373 collections as of November 25, 2008

American Memory.[26] This sampling approach allowed for detailed comparison of two different metadata sets (and approaches to collection-level description) for the same collections. The sample was further expanded through systematic sampling of every 5th American Memory record for collections not overlapping with Opening History (starting with the 3rd record). Collection records were selected from the "list all collections" page organized by collection title: http://memory.loc.gov/ammem/browse/ListAll. php?title=1.This resulted in a sample of 39 collection records.

- Systematic sampling of every 10th collection record in The European Library aggregation (starting with the 3rd record). Collection records were selected from the browse page http://search.theeuropeanlibrary.org/portal/en/collections_all.html. After exclusion of the records that represented catalogs or finding aids to physical collections rather than digital collections, the sample size totaled 27 collection records.

While the collection description schema used by The European Library includes 15 elements (9 required and 6 optional), its portal only displays the title and description. American Memory follows the same approach and displays to the end-user only the title and free-text *Description* field. Both The European Library and American Memory use remaining collection-level metadata "behind-the-scenes." The XML files with complete collection records were provided for this research by contacts at The European Library and American Memory aggregations: Sally Chambers (The European Library interoperability manager) and Christa Maher (American Memory metadata librarian).

[26] As of February 1, 2009, 14 collections that were part of Opening History were also part of American Memory.

Detailed manual content analysis of all collection-level records in the Opening History aggregation was conducted with the focus on fields for describing subject matter (*GEM Subjects, Subjects, Geographic Coverage,* and *Time Period*), subject-specific information in the free-text *Description* field and in the *Objects Represented* field. Patterns of application were observed for the above-listed fields. The following information was collected for comparison with the use of subject-related fields in the two other aggregations:

- Which specific kinds of information about the digital collection are encoded in the free-text *Description* field?
- How does the information provided in free-text *Description* and four subject-specific collection metadata fields relate to each other (e.g., one-way or two-way complementarity, redundancy etc.)?

- What is the overall richness of collection-level metadata records?

In the Opening History sample, *Description* fields of 22 collection records that originate from Illinois Harvest project included the identical statement:

"Illinois Harvest is a free public gateway combining search, aggregation, and discovery services. We provide organized and thematic access to digitized and born-digital resources about Illinois, created by Illinois scholars, or included among the digital collections of The University of Illinois Library. The goal of the Illinois Harvest/Large-scale Digitization Initiative is to broaden our digital collections and to enhance access to those collections, as well as to complementary digital resources elsewhere."

The decision was made to exclude this part of the *Description* field from further analysis since it did not characterize the individual collections.

A sample of collection-level records in American Memory and The European Library was analyzed to determine:

- Which fields are intended for use[27] and are actually used for providing subject-specific information about digital collections?
- How are these fields used for describing subjects of collections?
- How does the subject-specific information provided in different fields relate to each other (e.g., one-way or two-way complementarity, redundancy etc.)?
- Which specific kinds of information about the digital collection are encoded in the free-text *Description* field?
- What is the overall richness of collection-level metadata records?

Manual content analysis is highly dependent on human labor to interpret categories in the text. The well-known limitation of this approach is the difficulty of achieving high consistency among coders (intercoder reliability). Such inconsistency is primarily due to the lack of agreement among people on the interpretation of *categories* that describe abstract concepts in the textual content. To address this limitation, a detailed coding manual was developed (Appendix D).

3.3 Phase 2. User Interaction Analysis

3.3.1 Transaction Logs

Transaction log analysis is one of the methods actively used for unobtrusive observation of user behavior in information retrieval systems and on the Web. The method relies on transaction monitoring software which automatically records interactions with the web server in an electronic file called a *transaction log*. The transaction logs usually record: number of page hits,

[27] Based on collection-level metadata schema and/or other aggregation documentation if available

number and types of files downloaded, referral URLs[28], Web browser used, query logs, date and time of the transaction, IP address or Internet domain of the user, and — in systems where user authentication is required — user IDs. Often transaction log file data are exported to other tools (e.g., spreadsheet software) for further analysis and use.

Transaction log analysis, defined by Peters (1993) as "the study of electronically recorded interactions between online information retrieval systems and the persons who search for the information found in those systems", evolved as a research method in late 1970s-early 1980s. It is considered to be an effective way to study such user activities as frequency and sequence of feature use, hit rates, error rates, user actions to recover from errors, session lengths, and to detect discrepancies between what users say and think they do and what they actually do when interacting with information retrieval system. In particular, transaction log analysis data is widely used by digital resource developers to identify user communities[29] and patterns of use, inform digital collection development decisions and redesign/development of digital library websites, and assess impact of providing additional digital content and redesign/redevelopment of websites on use (Covey, 2002).

In a review of methodologies and techniques for transaction log analysis, Jansen (2006) identified three stages of the transaction log analysis research process: collecting log data, preparing the data, and analyzing the prepared data, where the first two stages influence the third stage and the results significantly. Jansen & Pooch (2001) pointed out the lack of standardization in terminology and methods in Web user studies. The nature of Web searching has made conducting observations of individual Web users much more complicated compared to similar

[28] i.e., how users get to the website

[29] Static IP addresses and Internet domain information are used to identify broad user communities (e.g., in-the-library, outside-the-library, on-campus, off-campus, international, etc.)

observations of earlier user studies in OPACs and other traditional IR systems. Extracting and interpreting the data requires discussion and definition. For example, defining what constitutes an individual user session in an unauthenticated Web environment is one of the major difficulties of transaction log analysis (Covey, 2002). The methodology for conducting transaction log analysis, including detailed explanation of all three stages and the steps researchers need to take in web log data collection, preparation and analysis, are outlined in Jansen (2008).

The "search log analysis" (SLA) — a type of transaction log analysis which focuses solely on searching behavior, has been a widely used research method. For example, Jansen, Spink, & Pederson (2004) compared characteristics exhibited by searches (e.g., query length and Boolean usage rates) in different digital collections. Several recent transaction log analysis studies categorized subject searches on the Web into FRBR-like categories such as people, places, and things[30] (e.g., Spink et al., 2002; Koshman et al., 2006; Beitzel et al., 2007; Jansen, Spink, & Koshman, 2007). Some studies comparing Web and online catalog searching patterns (Jansen & Pooch, 2001) have been conducted. A number of log analysis studies have matched user search terms with controlled vocabulary terms and reproduced user queries in databases (e.g., Greenberg, 2001; Gault, Shultz, & Davies, 2002; Gross & Taylor, 2005; Nowick & Mering, 2003).

As part of this dissertation research, transaction log analysis was used to seek answers to these research questions:

How do scholarly users of cultural heritage aggregations approach collection-level information discovery?

[30] Roughly corresponds to FRBR *person*, *place*, and *object.*

How does collection-level user search data fit the FRBR model?

And, more specifically:

- What are the typical collection-level user interaction patterns in the Opening History aggregation, for example:
    - approaches used (e.g., browse, advanced search, basic search, etc.)
    - use of search limits or Boolean operators
    - query length and frequency?
- How do collection-level and item-level user interaction characteristics compare?
- What is the distribution of search categories in collection-level user searching in Opening History aggregation?
    - What (if any) are the categories not covered by FRBR model and previous analysis?
- How do the distributions of the FRBR-based search categories compare at collection-level and item-level?

The analysis was conducted both at the collection-level and — as a point of comparison — at the item-level, and specifically focused on individual user queries rather than complete sessions. Due to the time constraints and the limitations of the transaction logging tools used by Opening History aggregation at the time of data collection (Google Analytics), the session-level analysis was left out of the scope of this study.

The systematic sampling approach was used. Data was drawn from a sample of user interactions with Opening History and/or its predecessor DCC aggregation. One week of user interactions was drawn from each of the 12 months between February 2008 and January 2009.

This resulted in a sample of 12 weeks of log data spread evenly throughout the year: February 8-14, March 15-21, April 22-28, May 1-7, June 8-14, July 15-21, August 22-28, September 1-7, October 8-14, November 15-21, December 22-28 of 2008, and January 1-7 of 2009. This approach helped capture seasonal variations in search behaviors and contributes to the reliability and generalizability of the results.

After filtering out agent queries performed by web spiders and queries initiated by DCC project staff from transaction log files, the transaction log data collected by the Google Analytics application was imported into MS Excel spreadsheet files and arranged into several groupings for further analysis:

- collection search
- collection browse, including browse by:
  - collection title
  - hosting institution
  - subject
  - geographic coverage
  - object type
- viewing of collection-level metadata records
- item search, including:
  - simple search
  - advanced search
- item browse, and
- other (spam, retrieval of cached previous search results, international searchers' queries made from Google language translation page, etc.).

Collection and item search queries were grouped with identical queries. This resulted in 501 unique collection search queries and 713 unique item search queries. Preserving the context of a search was considered an important factor for categorizing searches. Therefore, the decision was made not to parse queries into separate words or even further — into stems. Minimal processing of the queries was undertaken: both correct and misspelled versions of the same words (e.g., "immigration" and "imigration") were considered the instances of the same unique search query.

The research procedure used in transaction log analysis was tested in a pilot study (Zavalina, 2007) and included:

- Measuring query length and query frequency using traditional definitions and approaches.[31]
- Categorizing unique search queries into 11 FRBR-based search categories, including 7 FRBR entities (*work*[32], [individual] *person*, *corporate body*, *concept*, *object*, *event*, and *place*), 1 FRAD entity (*family*), 2 categories derived from the pilot study (*class of persons*, and *ethnic group*), and *unknown* search category.[33]

Coding of the user keyword searches was based on the procedure described by the Coding Manual (Appendix C). As with any categorization, this approach is inevitably subjective.

[31] See, for example Spink, Wolfram, Jansen, & Saracevic (2001). Query length — number of words in a query. Query frequency — number of times query used in a log.

[32]The FRBR *expression, manifestation* and *item* entities have not been adopted as categories for this analysis — although the cataloging has been traditionally performed for the manifestation level, it is virtually impossible to detect from the transaction log data alone what exactly the user is searching for: an abstract work, its particular expression, manifestation or item. Therefore, in the classification of the collection queries adopted for this study, *work* category is broader than FRBR *work* and covers any intellectual or artistic creation that has a title attribute, including the digital *collections* that are members of the Opening History aggregation.

[33] Unknown search category includes indiscernible search terms. Non-English search terms were categorized in appropriate category together with their English-language counterparts (e.g., tramvia and tram) whenever possible, otherwise placed in the unknown category.

This subjectivity constitutes one of the limitations of this study. The reliability of interpretation was increased through consultations between the principal investigator and a group of other researchers. After the list of coding categories had been revised by principal investigator based on initial coding, a sample of the coding was reviewed with a group of Metadata Roundtable members at the University of Illinois in December 2009.

Another limitation of subject search categorization lies in the ambiguity and polysemy of the actual queries. In order to minimize this limitation several research design features were included:

- Detailed coding guidelines assisted coders and allowed them to assign ambiguous user searches to multiple categories. As a result, collection search queries were on average assigned to 1.3992 categories, and item search query — to 1.3955 categories.
- Search categorization analysis results were triangulated with user interview and observation data.

The most widely acknowledged limitation of transaction log analysis is its inability to provide adequate data on "why users searched in the way they did". User needs, thoughts, goals and emotions at the time of the transaction are not reflected in the log data. The usual recommendation is to supplement transaction log analysis with other obtrusive and/or unobtrusive research methods such as questionnaires, protocol analysis, and interviews. To compensate for the shortcomings of transaction log analysis, interviews and observations of users searching aggregations of digital collections were incorporated into this study.

In addition, all the collection-level user searches obtained from the transaction logs based on the sampling technique described above were replicated in the Opening History aggregation in order to collect information about the specific fields in collection-level metadata where the match to the user keyword query appears and how often the query is satisfied only through the free-text collection metadata, or only through any of the subject-specific collection metadata fields. This investigation sought the answer to the question: ***Which fields of collection-level records provide scholarly users with the most valuable information to meet their needs?***

To answer this question, a method applied for this part of the analysis was developed by Gross and Taylor (2005), adapted for the purposes of this dissertation research. Using captured searches from the previously described transaction log data, a series of keyword searches were performed in the Opening History to determine what proportion of the collection records retrieved by each user's search had a keyword only in the free-text *Description* or other subject-specific collection metadata fields, and thus would not have been retrieved if the free-text collection metadata or subject headings were not there. For each term or set of terms, the number of hits with all keyword(s) anywhere in the collection record(s) was recorded, including the following kinds of data:

- number of collection records with at least one keyword in any collection metadata field(s),
- number of collection records with all keywords located in only one collection metadata field.

3.3.2 Interview and Observation Sessions

Interview and observation methods were used in Phase 2 of this research to collect real-life data to triangulate and clarify the findings of the transaction log analysis and content analysis and to contribute to answering the question:

How does collection-level metadata facilitate scholarly access? and its more specific parts:

- How do scholarly users of cultural heritage aggregations approach collection-level information discovery?
- Which collection-level metadata fields provide scholarly users with the most valuable information to meet their needs?
- How does collection-level user search data fit the FRBR conceptual model of subject entities?

Semi-structured interviews — alone or in combination with observations — have been widely used to study scholars' (and particularly, humanities and social science scholars) information seeking, including searching and browsing. Scholars are one of the major audiences for the Opening History and other similar aggregations of digital collections: for example, 54% of the respondents to an IMLS survey (*Assessment of End-User Needs in IMLS-Funded Digitization Projects*, 2003) name scholars as a target audience, while 88% of the respondents to a 2006 survey of DCC collections name scholars one of their target audiences (Palmer, Zavalina, & Mustafoff, 2007). Moreover, the Opening History aggregation specifically targets academic historians and history enthusiasts as a user group.

A small sample of the scholarly end-users was developed based on the subject strengths of Opening History and American Memory aggregation. This included topics such as Midwest and particularly Illinois history, American South history, Native American history, African American history, Japanese American history, etc. Local University of Illinois faculty and PhD Candidates in the Department of History who are actively publishing about history of U.S. states and regions, as well as other topics widely represented by collections in Opening History were targeted and recruited for one-on-one interview and observation sessions, based on their willingness to be interviewed.

Analysis of Department of History website[34] and the Illinois IDEALS institutional repository[35] resulted in a pool of 43 potential interviewees consisting of core History Department faculty members and Doctoral Candidates with diverse research interests, including 19th century U.S. political history, African American history, Civil War history, history of education in the U.S., Illinois and Midwest history, Latin American history, Native American history, U.S. economic and consumer history, U.S. immigration history, U.S. race and class history, U.S. South and Caribbean history, and U.S. women's history. Preference was given to scholars either currently involved in research on an aspect of U.S. history or those who have recently finished a major research project (dissertations defended and books/chapters or articles published within the last three years, 2007-2009). Six historians with areas of research that closely correlated with subject strengths of American Memory and Opening History were invited to participate in the interviews. Three of them participated in interview and observation sessions in February-April 2010.

[34] http://www.history.uiuc.edu
[35] http://www.ideals.uiuc.edu

To obtain information about the user-perceived role of collection metadata in subject access and resource discovery, interviews with these historians were combined with observation. Two U.S.-based aggregations, which use different approaches to displaying collection-level metadata were offered to the participants: 1) American Memory, that displays only the free-text collection metadata (*Description* field) to the end-user and uses the remaining rich collection metadata fields and values "behind the scenes" to support information retrieval and 2) Opening History, that displays collection-level metadata records to the user in their entirety. Historians were asked to explore each aggregation for content relating to a topic of their research and/or teaching. Historians were asked to compare their experiences in these two different environments, with data collected through a semi-structured interviews protocol. The average interview/observation session duration was 45 minutes. The interview/observation guide is attached (Appendix A). This interview strategy added richer data to the more generalizable quantitative results.

Since Opening History is a new aggregation of cultural heritage digital collections, created only in 2008, it was logical to assume the low awareness of this aggregation among the target audience — historians. Therefore interview/observation sessions were crafted so that information about both actual and potential use of Opening History aggregation was collected and assessed. On the other hand, American Memory is a long-standing and well-known aggregation with a focus on United States history, so the expectation was that at least some of the respondents would have prior experience interacting with American Memory.

Respondents with prior experience using Opening History and/or American Memory aggregation were asked to answer a number of questions regarding their most recent interaction(s) (critical incidence technique). One limitation of this strategy is the fact that it relies

heavily on the recollection of past events, thus the data collected might be limited by decreasing precision in recollection of events of variable currency. To account for such influence, the respondents were asked about the (approximate) date of the search they recounted. Interview data collected this way was combined with observation of participants' interactions with the Opening History and American Memory aggregations. In this way, respondents could support their recollections by repeating their searches while being observed.

Respondents with no prior experience using Opening History and/or American Memory before were given time to become familiar with the aggregation(s), asked to think about the topic(s) related to their current or recent research, and then asked to use Opening History and American Memory to locate objects covering these topics.

Each interview/observation was audiotaped and fully transcribed. User interactions with Opening History and American Memory aggregations conducted during the observation sessions were recorded with the help of *Camtasia Studio Screen Recorder and Presentation Software*. Interview/observation data were managed and analyzed qualitatively through exploratory data coding. The types of interview questions asked and the types of questions answered were coded by principal investigator, who also conceptualized and integrated data in an effort to explain its meaning, and identified features of information behavior in order to characterize the information seeking patterns of scholarly historians. The coding categories that emerged in this analysis included:

- familiarity with aggregation,
- interactions with aggregation:
    - viewing collection metadata records
    - viewing item metadata records,

    - collection browse (subject browse, geographic browse, project browse, object type browse, institution browse, and collection title browse)
    - item browse
    - collection search (basic and advanced)
    - item search (basic and advanced)
    - clickthrough to collection homepage
    - use of controlled vocabulary terms and Boolean operators
- differences in use of aggregation for research and teaching
- value of collection metadata to historians
- information and collection metadata fields important for the scholarly users
- reaction to collection metadata display, and
- other considerations (e.g., digitization practices, search results ranking, reactions to item-level metadata quality, values in collection metadata fields, etc.).

The research design included the use of three research methods — content analysis, transaction log analysis, and interview/observation — in order to significantly increase the reliability of the results by triangulation of data sources. The interview and observation data was used as secondary data to document real-life searching situations and the collection-level information seeking patterns of scholarly historians. The findings from the content analysis (e.g., the structure of collection-level records in two aggregations, the fields intended for subject representation, the controlled vocabularies used, etc.) informed the interview/observation process. Taken together, the data collected through these three research methods provided a coherent picture of collection-level information discovery by scholarly historians in aggregations of digital content and the role played by collection-level subject metadata.

3.4 Tables

Phases of analysis	Methods of data collection and analysis	Research question(s) answered
1. Collection metadata analysis	1. Comparative content analysis of collection-level subject representation in three large-scale aggregations of digital collections	What is the variation in richness of collection-level subject metadata across collections and aggregations of digital collections?
2. User interaction analysis	2a. Transaction log analysis of collection-level user queries in an aggregation, analysis of collection records from user search result sets, comparison with item-level search queries	How does collection-level metadata facilitate scholarly access? • How do scholarly users of cultural heritage aggregations approach collection-level information discovery? • Which collection-level metadata fields provide scholarly users with the most valuable information to meet their needs? • How does collection-level user search data fit the FRBR model?
	2b. Interviews with and observations of scholarly historians using two kinds of cultural heritage aggregations: • with collection-level metadata entirely displayed to the end-user • with only *Description* metadata field displayed to the end-user.	How does collection-level metadata facilitate scholarly access? • How do scholarly users of cultural heritage aggregations approach collection-level information discovery? • Which collection-level metadata fields provide scholarly users with the most valuable information to meet their needs? • How does collection-level user search data fit the FRBR model?

Table 2. Research phases and methods

Chapter 4. Collection Metadata in Aggregations of Digital Collections: Findings and Discussion

The results presented here are based on a systematic, manual content analysis of the collection-level metadata records in the three aggregations of digital collections. Research aims were addressed by identifying patterns in the data provided in the free-text *Description* field and other free-text and controlled vocabulary fields providing subject-specific information.

4.1 Subject-Specific Collection Metadata Fields and Controlled Vocabularies

The collection metadata fields intended for providing subject-specific information about digital collections, including genre and object type information, in three aggregations are listed in Table 3. Table 3 demonstrates how mapping of these fields (which are named differently in the different aggregations) was achieved. The "common-denominator" field names in bold in the far-left column — *Subjects, Objects, Temporal Coverage, and Geographic Coverage* — are used to report results of the study, starting with the section 4.2 of this thesis.

Among the three aggregations studied in this research, only the Opening History displays its collection-level metadata in its entirety to the user, while the other two aggregations keep most of their collection-level metadata (except for the *Title* and free-text *Description* field) "behind-the-scenes."

4.1.1 Opening History

As discussed in section 1.6, the Opening History aggregation uses four subject-specific metadata attributes in its Dublin Core Collection Application Profile (DCCAP)-based collection description scheme:

- topical,
- geographic,
- temporal, and
- free-text description.

In addition to these elements, in the previous pilot studies, the *Objects Represented*[36] field was also found to be subject-rich, especially for describing genres of objects in a digital collection. Browsing in Opening History is supported through Gateway to Educational Materials (GEM)[37] subject categories, a vocabulary developed for educators and which the initial developers of the IMLS DCC considered suitable for browsing cultural heritage materials. It consists of twelve broad top-level subject headings (e.g., Arts; Social Studies), each with between twelve and twenty-nine narrower second-level headings (e.g., Architecture; State History). The use of at least one top-level GEM headings in the *GEM Subjects* field is required, and the use of second-level GEM headings is optional. Three additional optional fields are intended in DCC collection-level metadata scheme for subject access: *Subjects* (for controlled

[36] Opening History's *Objects Represented* field was mapped to the "common-denominator" *Objects* field for comparative content analysis of collection metadata in three aggregations (see Table 3).
[37] http://www.thegateway.org/about/documentation/gem-controlled-vocabularies/vocabulary-subject

vocabulary terms or uncontrolled keywords other than GEM), *Time Period*, and *Geographic Coverage*[38].

Library of Congress Subject Headings (LCSH) are often used for the *Subjects* field, although the use of LCSH terms is not formally required. The use of terms from Getty's Thesaurus for Geographic Names (TGN)[39] is strongly recommended in *Geographic Coverage* field. For describing *Time Period*, the online Collection Registry entry form (Figure 5) provides a checklist of date ranges suggested by Opening History developers (e.g., 1850-1899, 1930-1949), as well as a free-text field which allows expressing temporal coverage differently and often more specifically (e.g., "Civil War", "1939-1945"). Similarly, the required *Objects Represented* field allows using both controlled vocabulary values suggested by Opening History developers (e.g., Photographs/Slides/Negatives, Books and pamphlets) and alternative free-text values (e.g., "diaries", "aerial views"), while the use of the Library of Congress Thesaurus for Graphic Materials II: Genre and Physical Characteristic Terms (TGM II)[40] is encouraged.

4.1.2 The European Library

The European Library aggregation uses The European Library Application Profile for Collection Descriptions (ELAPCD[41]). Similarly to IMLS DCC collection metadata schema used in Opening History, ELAPCD is a Dublin Core-based application profile, which is based on the RSLP Collection Descriptions scheme[42] and has been updated to be largely compatible with the

[38] Both *GEM Subjects* and alternative *Subjects* were mapped to the "common-denominator" field *Subject* for comparative analysis of collection metadata in three aggregations. *Geographic Coverage* and *Time Period* fields were mapped to *Geographic Coverage* and *Temporal Coverage* "common-denominator" fields respectively (see Table 3).
[39] http://www.getty.edu/research/conducting_research/vocabularies/tgn
[40] http://www.loc.gov/rr/print/tgm2
[41] http://www.theeuropeanlibrary.org/handbook/Metadata/tel_ap_cld.html
[42] http://www.ukoln.ac.uk/metadata/rslp/schema/

NISO Collection Description Specification[43] and the Dublin Core Collection Description Application Profile.[44]

The ELAPCD scheme includes a total of four fields for subject-specific information: *Subject*, *Spatial Coverage*, *Temporal Coverage*, and *Theme* (a field which provides genre information).[45] Only *Subject* is a required field, while *Spatial Coverage*, *Temporal Coverage*, and *Theme* fields are optional. Additional required field — *Collection Item Type* — uses only two values ("digital" and "not digital"[46]) and therefore was excluded from the analysis of subject-specific metadata.

The European Library uses Universal Decimal Classification (UDC) subject terms and numeric notations (e.g., "760 Graphic arts", "940 General history of Europe") for topical subject headings. In cases of comprehensive digital collections the value "all subjects" is used in *Subjects* field instead. The European Library enforces the use of controlled vocabulary in *Subject* and *Theme* fields by providing dropdown menus with suggested values for these two fields in its Collection Description Editor entry form (Figure 6). There is no indication that any controlled vocabulary is required for use in *Spatial Coverage* and *Temporal Coverage* collection metadata fields.

4.1.3 American Memory

While the American Memory aggregation uses several metadata schemes for describing its collections — MARC21XML, MODS, and Dublin Core Collection Application Profile —

[43] http://www.niso.org/workrooms/mi/Z39-91-DSFTU.pdf
[44] http://dublincore.org/groups/collections/collection-application-profile/
[45] These ELAPCD fields were mapped to "common-denominator" fields − *Subject*, *Geographic Coverage*, *Temporal Coverage*, and *Object* — for comparative analysis of collection metadata in three aggregations (see Table 3).
[46] The European Library includes both digital and physical collections created by the national libraries in Europe.

only the MODS metadata set was provided by American Memory for this analysis. In the MODS collection metadata set, two major subject-specific fields are used: *Subject* and *Type of Resource.*[47] However, *Subject* fields are further subdivided into the more specific subfields or facets: *topic*, *geographic*, *temporal*, etc.[48] American Memory uses LCSH for topical subject headings, Library of Congress Name Authority Files (LC NAF) for personal, corporate, and geographic names, and Index Terms for Occupations in Archival and Manuscript Collections (ITOAMC) for occupation subjects. It is unclear is TGM II controlled vocabulary is used for *Type of Resource* and *Genre* fields.

4.2 Metadata Richness

The three aggregations vary in granularity of their subject-specific collection-level metadata. Figures 7, 8 and 9 show subject representation in MODS collection metadata in American Memory, and DCCAP-based collection metadata records in The European Library and Opening History. In The European Library, two free-text collection metadata fields — *Title* and *Description* — are presented in 28 different European languages (see for example Figure 10). Work is ongoing in The European Library to translate the entire metadata records into 27 languages other than English. This added level of complexity calls for more concise free-text *Description* fields, as well as other collection metadata fields in the aggregation, and somewhat reduces overall metadata richness discussed in this section.

[47] In a small proportion of collection metadata records in the AM sample, *Genre* field was also included. Both *Type of Resource* and *Genre* were mapped to "common-denominator" *Objects* for comparative analysis of collection metadata in three aggregations (see Table 3).

[48] These three fields were mapped to "common-denominator" *Subjects*, *Geographical Coverage*, and *Temporal Coverage* respectively for comparative analysis of collection metadata in three aggregations (see Table 3). Other, rarely used, *Subjects* subfields in the AM sample included *name* and *occupation*. These were mapped to "common-denominator" *Subjects*.

For the purposes of this research, metadata richness is defined as a combined measure of the following indicators:

1. the number of subject-specific collection-level metadata elements used in collection-metadata record,
2. the length of the free-text *Description*, *Subjects*, *Temporal Coverage* and *Geographic Coverage* field measured as the number of words per field,
3. the number of collection properties (e.g., uniqueness, provenance, subject, object, navigation and functionality, etc.) represented in the free-text collection-level metadata (e.g., *Description* field),
4. mutual complementarity of the values in collection metadata fields.

The following sections detail the findings of comparative analysis of collection metadata richness in three aggregations.

4.2.1 Application of Subject-Specific Metadata Fields

Free-text *Description* is a required collection metadata field in all three aggregations, and therefore is consistently applied in 100% of collection records in the sample. Figure 11 shows the frequency with which other subject-specific metadata fields are applied in the analyzed collection records across the three aggregations. Opening History consistently uses all four subject-specific metadata fields in 100% of its collection records,[49] while both American Memory and The European Library only consistency in apply collection *Subjects* and/or *Objects* fields. The subject-specific metadata field that is used the least consistently in American Memory

[49] Out of these subject-specific collection fields, only *GEM Subjects* and *Objects* fields are required field in DCC collection metadata schema that is used in Opening History. However, all four subject-specific fields are used in 100% of the records.

is *Temporal Coverage*, while The European Library pays the least attention to the use of the *Objects* field. This much higher consistency in application of subject-specific metadata fields in Opening History can be explained by the fact that creation of collection metadata records in this aggregation is centralized. While the content for collection metadata in Opening History is drawn directly from documentation provided by the local developers of the individual collections, collection records are created manually by the IMLS DCC project staff members who follow well-developed guidelines, and special attention is paid to the completeness of collection metadata records.

4.2.2 Length of Free-Text *Description* and Other Subject-Specific Metadata Fields

There is considerable variation in the length of (or number of words used in) the *Description* field both within each aggregation and across the three aggregations. Figure 12 displays the comparative frequency distribution of the *Description* field length values in three aggregations. American Memory *Description* fields are the longest, with 132 words on average. Opening History *Description* fields contain 98 words on average, while The European Library *Description* fields contain on average 45 words. In Opening History, *Description* fields exhibit the widest range in length: from 5 to 429 words. American Memory *Description* fields range in length from 32 to 260 words, while The European Library *Description* fields vary the least among the three aggregations: from 13 to 114 words.

Table 4 provides additional statistical information on the length of the free-text *Description* field in the three aggregations. It shows that while the mean and median lengths of the *Description* field are the lowest in European Library, the standard deviation is also the

lowest, which means the length of *Description* fields is more consistent in this aggregation. Both average and median values of the *Description* field length are the highest in American Memory, with higher variability (variance and standard deviation) than in the European Library. Opening History exhibits the highest variability of *Description* field length, with average and mean indicators higher than in The European Library and lower than in American Memory.

However, quantitative analysis of the other subject-related fields in collection metadata, as presented in Table 5, demonstrates that not all of the subject-specific collection metadata fields are the longest in American Memory. For example, Opening History has the longest *Geographic Coverage* fields, with the highest mean and median values. Two more metadata fields — *Subjects* and *Objects* — have the highest median length in Opening History, although American Memory has the highest average length. At the same time, The European Library has the highest median length for the *Temporal Coverage* field, although this field has the longest average length in American Memory.

As illustrated by Table 6, American Memory exhibits the highest cumulative length of *Description* and other subject-specific collection metadata fields among the three aggregations. Opening History uses somewhat more concise collection-level subject metadata, while The European Library uses the most concise collection metadata overall. However, it is important to note here that Opening History is the only one among the three aggregations that displays its collection-level metadata in its entirety to the user, while the other two aggregations keep most of their collection-level metadata (except for the free-text *Description* field) "behind the scenes."

4.2.3 Collection Properties in Free-Text *Description* Field

The list of collection properties discussed in this section was developed in a pilot study (Zavalina et al., 2008a, 2008b), working with a second coder to develop agreement on the categories and the terminology through iterative review and discussion. A total of nineteen collection properties were identified in the pilot study: subjects, object types/genres, geographic and temporal coverage, creators of items in collection, collection title, size, collection development policy, copyright information, provenance, collection's importance, uniqueness, comprehensiveness, intended audience, navigation and functionality, language of items in collection, frequency of additions to collection, participating/contributing institutions, and funding sources.

The collection properties data in this section is based on the detailed manual content analysis of all 554 free-text *Description* fields in the main study sample performed in the Summer-Fall 2009. The intercoder reliability test on a random sample of 6 *Description* fields (2 collection records from each of the 3 aggregations) was conducted with seven other coders — graduate students of the University of Illinois Graduate School of Library and Information Science — in Spring 2010. The coders were recruited from the Center for Information Research in Science and Scholarship (CIRSS) Student Research Group. The coding guidelines used for the test are included in Appendix D. The overall intercoder reliability was calculated at 90.02%. The highest intercoder reliability of 96-100% was recorded for the following collection characteristics encoded in the free-text *Description* field: copyright, frequency of additions, funding sources, language of items in collection, navigation and functionality, and size. The lowest intercoder reliability of 79-83% was recorded for the following collection characteristics:

provenance, geographic coverage, comprehensiveness, importance, and collection development information. The detailed matrix of intercoder reliability data is included in Appendix E.

Across the three aggregations, the average *Description* field provides information about 6 collection properties. All 19 collection properties were found in collection records in the Opening History, while American Memory and The European Library collection metadata samples contained 18 collection properties each. American Memory *Description* fields lack frequency of additions information, while The European Library *Description* fields lack funding sources information.

American Memory, which has the longest *Description* fields among the three aggregations, also exhibited the highest average and median numbers of collection properties encoded in *Description* fields, with between 1 and 12 collection properties (Table 7). While the medium positive correlation between the length of *Description* field and the number of collection properties was observed in all three aggregations, the highest Pearson R value (0.513432) was recorded in the American Memory. While indicating somewhat higher overall richness of *Description* fields in American Memory, this finding also suggests that the longer free-text *Description* fields tend to provide richer description of digital collections.

Figure 13 and Table 8 provide a side-by-side comparison of the distribution of collection properties found in the free-text *Description* fields in the three aggregations.

Subject information was the most widely represented collection property in the free-text *Description* field, with 95% of the collection records overall (100% in American Memory, 96% in Opening History, and 74% in European Library. The content ranges from very specific subject coverage statements (e.g., “cover a broad range of topics, including ranching, mining, land grants, anti-Chinese movements, crime on the border, and governmental issues”) to broader

subject coverage statements (e.g., "in the fields of culture, education, and academic research"), to subject keywords scattered throughout the text, as in this example: "During World War II, as a member of the *U. S. Army, 252nd Field Artillery Battalion*, he captured over 700 images of *life as a soldier* and unique snapshots of *events of the war*."

Object type information was the second most prominent in the free-text *Description* field, with 90% of the collection records describing types of digital objects in a collection (100% in American Memory, 89% in Opening History, 85% in European Library). General object terms, such as "physical artifacts," were common, as were more specific terms, such as "lanterns, torches, banners." Physical formats and genres are also frequently specified, as with "pamphlets, leaflets, and brochures", "songbooks", "political cartoons," and "chronics, letters, annals, official documents." Object types and formats are sometimes conflated, even within the same sentence, in the *Description* field, as well as in *Objects Represented*. This lack of disambiguation between object type and format is a known metadata quality problem for digital object description[50] (Jackson et al., 2008; Godby, Smith & Childress, 2003; Park, 2005; Hutt & Riley, 2005).

Geographic and temporal coverage of a digital collection were the third and fourth most widely represented collection properties in free-text *Description* fields in three aggregations. Geographic coverage information was found in 79% of collection metadata records overall (81% in Opening History, 69% in American Memory, and 56% in the European Library). Indications of geographic coverage of varying granularity (e.g., "Austro-Hungarian Empire", "Dutch Indies", "Mayan city of Uxmal in Yucatan, Mexico and a Native American Mississippian site, Angel Mounds U.S.A.") were found in free-text *Description* fields. Indications of temporal

[50] This problem is mentioned as an example of collection-level metadata patterns revealed by this study, however, the detailed discussion is out of scope of this study.

coverage in the *Description* fields were found in 65% of collection metadata records overall (85% in American Memory, 65% in Opening History, and 48% in the European Library). These indications ranged from specific dates and date ranges (e.g., "19th century", "covering the period of 1894-1932, with the exception of 1896"), to known historical periods (e.g., "World War I", "California Golden Rush"), and finally to combinations of the former two approaches (e.g., "Lithuanian press ban period (1864-1904)."

Names of artists or institutions that created items in the collection were found in 42% of *Description* fields overall (42% in Opening History, 41% in American Memory, and 33% in European Library. For example, corporate authors may be identified as in "The Museum Extension Projects of Pennsylvania, New Jersey, Connecticut, Illinois, and Kansas crafted most of the items currently in the collection", or "monasteries of Mount Athos: Chilandar, Vatoped, Simonopetra and Kutlumush." Individuals might be specified (e.g., "Among the authors represented are Frederick Douglass, Booker T. Washington, Ida B. Wells-Barnett, Benjamin W. Arnett, Alexander Crummel, and Emanuel Love") and further biographical information for them supplied as well (e.g., "images are noted on their mounts as being from Watkins's 'New Series'.... Watkins was active between 1854 and the late 1890s."). There is no specialized element in the collection metadata schema used by any of the three aggregations that could accommodate this type of information,[51] yet it appears of high value as contextual information for users. The collections related to single or multiple authors could benefit from more formal representation of item creators. In this case, a new element would need to be specified, since the

[51] Opening History aggregation's collection description metadata scheme currently uses dc:creator element in a limited way, to indicate a grant project responsible for creation of digital collection, but overlooks creators of physical items and physical collections.

existing Dublin Core Collection Application Profile (DCCAP) *Collector* element is designed to cover the creator of the collection, not the creator of items in the digital collection.

Forty-two percent of the free-text *Description* fields (56% in American Memory, 48% in European Library, and 30% in Opening History) contain either explicit or implicit evidence of certain collection development policies, or digitization selection guidelines. Some of the more specific descriptions offer information such as: "titles published between 1850 and 1950 were selected and ranked by teams of scholars", "effort has been made to offer a balanced number of items for each inaugural event", to more ambiguous criteria, as in: "a selection of framed items from the collections of the ... Library," or "a sample of the photographic archives." Some descriptions identify plans for future collection development, a potentially significant aspect of collector intentionality, or other locally accessible assets: "in addition to the newspapers, it is planned to provide access to a complimentary collection of Richmond related Civil War period resources", "additional lesson plans, activities and photo essays designed by teacher advisors and educational consultants will be added in the future." Other free-text *Description* fields explicitly state a purpose: "stimulate the documentation and preservation of ethnic materials and foster a greater interest in the history and cultures of the peoples of the region", "to inventory and to describe the decoration of the manuscripts held in the Bibliothèque Nationale de France." These statements are multifaceted, with important data about potential audiences and the intellectual and evidentiary intentions of collectors.

Collection title information — complete title, subtitle, or title acronym — was the seventh most often found collection property in free-text *Description* fields across the three aggregations (41% overall: 72% in American Memory, 39% in Opening History, and 30% in European Library). Many titles provide concise statements with subject-specific information

(e.g., “The 1936 Gainesville Tornado: Disaster and Recovery”), as well as information on the types of objects in the collections (“Warsaw in Words and Images”), which are typical of the *Description* field content.

Provenance information was included in 36% of the free-text *Description* fields overall (33% in American Memory, 36% in Opening History, and 37% in European Library. These sample excerpts represent the kinds of information provided: “in December 2002, the ... Library acquired the Humphrey Winterton Collection of East African photographs, “acquisition of these hitherto unknown manuscripts was spearheaded by Edgar J. Goodspeed in the first half of the twentieth century”, “a 1988 bequest of more than 850 landscape prints and drawings from the collection of Los Angeles architect Rudolf L. Baumfeld significantly enhanced this wide-ranging and well-studied thematic area”, “selected from various Library of Congress holdings”, “documents belonging to the collection of the Army Museum.”

The three aggregations include a large number of museum, historical society or archive collections. It seems likely that a provenance element might serve even a greater percentage of collections than those who exploit the *Description* field for this purpose. The Dublin Core Collection Description Application Profile’s *Custodial History* element covers provenance information found in the free-text metadata. However, at the time of analysis, this element was not used by any of the three aggregations. The interviews and observations of scholarly historians, performed as part of this dissertation research, have shown that the scholars consider the presence of provenance information in collection metadata as crucial and would prefer to see it in a specialized field.

A third of collection records (90% in American Memory, 37% in European Library, and 28% in Opening History) had *Description* fields that made statements about the collection size, ranging from quantitative specifications ("175 engravings", "more than 70,000 volumes of digitized texts, 80,000 still images, and 30 hours of sound recordings", "28,000 of primary-source items") to general orientations (e.g., "hundreds of personal letters, diaries, photos, and maps"). Some free-text *Description* fields also referred to the size of an associated physical collection, such as: "the costume collection at the ... Museum has over 30,000 items of clothing and accessories" "the physical collection contains over 400 garments", "physical collection is comprised of several hundred photographs, publications and newspaper clippings," etc.

Fifteen percent of *Description* fields (21% in American Memory, 14% in Opening History, and 11% in European Library) contained navigation or functionality information (e.g., "may be searched or browsed in a variety of ways, including by keyword, subject, creator, title, and date", "allows the user to browse the highlights thematically or by number", "accessed by the scanned county photomosaic or line indexes", "accessible by date of issue or by keyword searching," etc.). This excerpt shows the kind of functions associated with a collection of television programs: "video excerpts, searchable transcripts, a select number of complete interviews for purchase, and resource management tools." Some of the statements in the *Description* field were accompanied by information on how the digital collection is organized for browsing (e.g., "grouped by county", "the overall organization of the database is by tribe", "arranged chronologically by Japanese periods", "organized according to seven major categories. Because a map will be assigned to only one category, unless it is part of more than one core collection, searching the Collection at this level will provide the most complete results since the indexes for all categories are searched simultaneously" etc.).

Overall, 13% of free-text *Description* fields (46% in American Memory, 15% in European Library, and 10% in Opening History) provide information about institution(s) hosting the digital collection, participating in the digitization project, and/or contributing items to digitize Representative examples include: "project brings Tufts, and the Virginia Center for Digital History together with the University to build a digital repository", "digital images of archival collections located at three Arizona repositories: the University of Arizona Library Special Collections, the Arizona Historical Society-Tucson, and the Arizona State Library, Archives, and Public Records", "the … Archives Department provides access to the digitized Roman Catholic Church registers of birth, marriage and death (1599-1907). The … Art Museum presents digital images," etc.

Audience information, both broad and specific (and sometimes implicit), was found in 7% of *Description* fields overall (7% in Opening History, 7% in European Library, and 3% in American Memory). Representative examples of specific audiences listed include: "Alabama residents and students, researchers, and the general public in other states and countries", "created especially for middle and high school students", "for those studying political reorganization in Georgia and the growth of Atlanta as well as the Civil Rights Movement, the Cold War, the Vietnam conflict, Middle East tensions, and Watergate", "schoolchildren, genealogists, historians, authors, producers, and special interest groups." Examples of the implied general public and educator audience include: "provided for personal use or educational presentations", "used strictly for private purposes," and "made available to the public for the first time."

Five percent of free-text *Description* collection metadata fields overall (18% in American Memory and 4% in Opening History) acknowledge funding sources — public or corporate — that helped build the digital collection (e.g., "received an IMLS National Leadership grant to

create the digital resource", "funds provided by the Institute of Museum and Library Services, under the federal Library Services and Technology Act", "digitized as the result of an Illinois State Library FY98 Educate and Automate grant", "made possible by a major gift from Citigroup Foundation", "funded by Reuters America, Inc., and The Reuters Foundation", "made possible through the generous support of the AT&T Foundation" etc.). No indication of funding sources in *Description* field was found in a sample of The European Library collection metadata records.

Special claims about collections — importance, comprehensiveness, and uniqueness — are found in a limited number of *Description* fields (4%, 5%, and 6% overall respectively), but they are of particular interest as the kind of self-assessed, special claims used to distinguish special collections in libraries, museums, and archives. The discussion of specific occurrences will follow. These findings on special claims that developers make about their collections will not be surprising to the metadata community. For example, there has been discussion about the inclusion of a *Strength* element into the Dublin Core Collection Application Profile (DCCAP) to accommodate descriptive information related to aspects such as importance, uniqueness, and comprehensiveness (e.g., Johnston, 2003), while the RSLP collection description schema has an "cld:strength" element for "An indication (free text or formalized) of the strength(s) of the collection"[52] (e.g., Heery & Patel, 2000).

Indications of importance or significance of a digital collection (e.g., "collection of the most important and influential 19th and early 20th century American cookbooks", "materials are significant in their place within the fabric of American history and culture", "creating an archive of unparalleled importance", "important and rare books, government documents, manuscripts, maps, musical scores, plays, films, and recordings", "the most outstanding representatives of

[52] See http://www.ukoln.ac.uk/metadata/rslp/schema/

Yiddish literature," etc.) were found in 8% of *Description* fields in American Memory, 4% in Opening History, and 4% in European Library.

Indications of comprehensiveness, definitiveness, or richness were found in 8% of *Description* fields in American Memory, 7% in European Library, and 4% in Opening History. Examples include: "a comprehensive and integrated collection of sources and resources on the history and topography of London", "the most comprehensive library of manuscripts, rare and contemporary books", "one of the most ambitious and comprehensive effort to date to deliver educational content on the Civil Rights Movement", "such a large body of materials presents a full spectrum of representation and opinion", "a rich diversity of materials", "almost complete collection of Norwegian printed newspapers," etc.).

The indications of uniqueness or rarity of the content of a digital collection were found in a somewhat larger proportion of *Description* fields (18% in American Memory, 5% in Opening History, and 4% in European Library) than the other two special claims — comprehensiveness and importance. Examples include "unique historical treasures from ... archives, libraries, museums, and other repositories", "rare historic published monographs and serials", "rare and unique library and archival resources on race relations", "sources that are rare, unusual, out-of-print, or difficult, if not impossible, to access," etc.

Language of items in a digital collection was mentioned in 4% of *Description* fields overall (8% in American Memory, 4% in Opening History, and 4% in European Library). Representative examples include: "many of the publications are in Vietnamese", "English- and Yiddish-language playscripts", "European, Slavic, Middle Eastern, and English- and Spanish-

language folk music in one region of the United States", "a listing of faculty, officers and graduates, entirely printed in Latin."

The information about copyright and frequency of additions to the digital collection were the lowest across the three aggregations. Copyright information (e.g., "these materials are royalty-free and available free of charge", "materials with expired copyrights", "restricted to items that are not covered by copyright protection", "historical sheet music registered for copyright," etc.) was found in only 1% of *Description* fields overall (5% in American Memory, 4% in European Library, and 1% in Opening History). Either specific or vague, indications of the frequency of additions to a digital collection (e.g., "regular additions to the collection are expected", "some 10,000 volumes per year", "annual growth is ca. 700 publications," etc.) were found in 1% of *Description* fields in The European Library and 1% in Opening History, while no such indications were found in the sample of collection metadata records from the American Memory aggregation.

Differences, sometimes significant, in the frequency of the use of certain collection properties were observed among the three aggregations. Overall, 14 out of 19 collection properties were found more often in free-text *Description* fields in American Memory than in the two other aggregations, with the most pronounced difference in application of title, size, and hosting/contributing institutions properties. Special claims about digital collections — uniqueness, importance, and comprehensiveness — were also more consistently represented in American Memory. Two collection properties — provenance and frequency of additions — occur more often in The European Library aggregation. Two collection properties — geographic coverage and creator of items in a digital collection — were found more often in Opening History, while one more — intended audience or uses of a digital collection — was found

equally often in Opening History and The European Library and more often than in American Memory.

These differences might be explained by the specifics of the policies followed, and/or tools used in describing digital collections in three aggregations as well as collection development policies. For example, the fact that only free-text *Description* fields are displayed to the end-user in American Memory, might be influencing the decisions on how rich *Description* fields should be in this aggregation, and resulting in longer and richer *Description* fields. More consistent indication of uniqueness, importance, and comprehensiveness of a digital collection in *Description* fields may be due to American Memory's collection development policy, which emphasized digitizing collections of great value to historians. Provenance information might be emphasized in The European Library due to its context of a multi-national aggregation, where tracing collections' provenance becomes complicated. Wide-spread use of geographic coverage information in free-text collection *Description* fields of Opening History might be due to the focus on local history in its collection development policy.

4.2.4 Mutual Complementarity of Collection Metadata Fields

A significant proportion of collection records in the sample includes cases of one-way complementarity, when information in one field (most often, *Description*) complements information in other field, by providing additional details absent elsewhere. In 97% of records in the sample (100% in American Memory, 97% in Opening History, and 84% in The European Library), it is the *Description* field that complements information found in one or more of the fields intended for subject indexing*: Subjects*, *Geographic Coverage, Temporal Coverage,* and *Objects.*

As seen in the Figure 14, the free-text *Description* field most often (86% in American Memory, 76% in Opening History, 70% in European Library) complements topical information found in the *Subjects* field with essential subject information. Representative examples include: "Spanish cartographer", "history, urbanism, public works and agriculture from a strictly geographic point of view" in *Description* vs. "900 History and geography", "911 Historical geography" in *Subjects*; "interior design", "homes of U.S. presidents" in *Description*, with these topics not mentioned in *Subjects*; "early developments in the National Park", "landscape and park facilities" in *Description* vs. "Great Basin", "Social studies", "State history" in *Subjects*.

Temporal Coverage field is the second most often complemented by *Description* field (67% in Opening History, 51% in American Memory, and 15% in European Library). Representative examples include: "16th century, 17th century, 18th century, 19th century, 20th century" in *Temporal Coverage* vs. "Since the Eighty Years' War" in *Description*; "from 1895-1920s" in *Description* vs. "1850-1899, 1900-1929" in *Temporal Coverage* field.

Objects field is also often complemented by object-type or genre-specific information in *Description* field (70% in American Memory, 44% in European Library, and 30% in Opening History), as illustrated by the Figure 15. Representative examples include: "uniform books, ego documents, photographs and sketches" in *Description* vs. "images" in *Objects* ; "digital pre-print originals and online publications" in *Description* while *Objects* field is missing; "historical photographs", "portraits", "aerial shots" in *Description* vs. "photographs/slides/negatives" in *Objects*; "rare books, government documents, manuscripts, maps, musical scores, plays, films, and recordings" in *Description* vs. "software, multimedia" in *Objects*.

Geographic Coverage is complemented by the *Description* field the least often (39% in Opening History, 33% in European Library, and 19% in American Memory). Representative examples include: "Hispanic America", "Spanish territories in America and Oceania" in *Description* vs. "Hispanic America" in *Geographic Coverage*; "Hungary or the Central European region" in *Description* vs. machine-readable "hu" in *Geographic Coverage*; "American states, the District of Columbia, and London, England" in *Description* vs. "United States" in *Geographic Coverage*; "Baja California, Mexico in an area south-east of Ensenada" vs. "Mexico (nation)" in *Geographic Coverage.*

The cases of *Description* field complementing several subject-specific fields in the same collection metadata record were observed. In the example from the Opening History aggregation (Figure 16), the free-text *Description* includes keywords (e.g., "children's lore, foodways, religious traditions, Native American culture, maritime traditions, ethnic folk culture, material culture, and occupational lore") complement both *Subjects* and *Objects* fields with topical and genre information. The standard subject vocabulary options are clearly too general and the free-text description is, as one would expect, likely to be more compelling to users.

In addition to complementing information encoded in other subject-specific collection metadata fields (*Subjects*, *Objects*, *Geographic Coverage*, and *Temporal Coverage*), the *Description* field also often complements and clarifies values in other collection metadata fields. One of the examples of such complementarity, illustrated by the Figure 17, includes *Audience* field.

As illustrated by Figure 18, however, it is not only the *Description* field that complements information found in other fields. The analysis also shows that information

encoded in other subject-specific collection metadata fields often complements information in the *Description* fields.

The *Subject* field was found to complement information found in the *Description* field in 52% of collection metadata records (70% in Opening History, 60% in European Library, and 30% in American Memory). Representative examples include: "860 Spanish and Portuguese literatures" in *Subjects* with no mention of this topic in *Description*; Tennessee Valley Authority", "African Americans", "forestry" in *Subjects* with no mention of these topics in *Description;* "Chesapeake Bay Region (Md. and Va.)", "Washington Region" in *Subjects* vs. "Chesapeake Bay Region" in *Description*; 15 specific subject strings (e.g., "North Carolina — African-Americans, North Carolina — Agriculture, North Carolina — Economics and Business" in *Subjects* vs. "North Carolina", "story of the Tar Heel State" in *Description.*

The *Temporal Coverage* field was found to complement *Description* field in 43% of collection metadata records (72 % in European Library, 67% in Opening History, and only 3% in American Memory[53]). Representative examples include: "1400s-1699, 1700-1799, 1800-1849, 1850-1899, 1900-1929, 1930-1949, 1950-1969, 1970-1999, 2000 to present, Pre-1400" in *Temporal Coverage* while no time information is provided by *Description* field; "1783-1789"in *Temporal Coverage* while no information is provided by *Description*; "1200-1900" in *Temporal Coverage* vs. "European age of chivalry" in *Description.*

The *Geographic Coverage* field was found to complement *Description* much more often than the *Description* field complements *Geographic Coverage* field, or in 43% of collection metadata records (56% in European Library, 55% in Opening History, and 24% in American

[53] Low level of complementarity in the case of American Memory is explained by the fact that the *Temporal Coverage* collection metadata field is applied inconsistently in this aggregation.

Memory). Representative examples include: “Poland, Lithuania, Ukraine, Belarus” in *Geographic Coverage* vs. “Poland” in *Description*; “Germany” in *Geographic Coverage*, while no geographic information is provided in *Description*; “Europe”, “Italy,” Great Britain” in *Geographic Coverage* vs. “US and abroad” in *Description*; “United States (nation), Midwest U.S. (general region), Illinois (state), Randolph (county), Knox (county)” in *Geographic Coverage* vs. “Randolph County, Illinois” in *Description*.

The *Objects* field also often complements information found in *Description* field in two aggregations — Opening History (52%) and American Memory (14%) — for 23% of analyzed collection records overall. No such trend was observed in The European Library, which can be explained by inconsistent application of *Objects* field in this aggregation: in 59% of collection records in The European Library sample the *Objects* field is blank or missing, while in the remaining 41% this field contains one-word-long broad terms (e.g., “images”, “maps”). Representative examples of the *Objects* field complementing *Description* field include: “Film transparencies—Color”, “Cityscape photographs” in *Objects* vs. “photographs” in *Description*; “Gelatin silver prints”, “Safety film negatives”, “Nitrate negatives” in *Objects* vs. “original negatives and photographic prints” in *Description*; “books and pamphlets, photographs / slides / negatives, newspapers, posters and broadsides, periodicals, prints and drawings” in *Objects* vs. “manuscripts, photographs, ephemera and published materials” in *Description*.

The cases of two-way mutual complementarity between the two metadata fields are much less numerous than one-way complementarity (12% of collection metadata records in the sample). Over a half of these cases were found in The European Library aggregation. Mutual complemenarity was observed the most often between *Description* and *Subjects* fields and

between *Description* and *Geographic Coverage* fields, while the least mutual complementarity was observed between *Description* and *Objects* fields. Representative examples include:

- "Letters" in *Description v*s. "autograph albums" in *Subjects* (taken together, the values in two fields provide more comprehensive genre information).
- *"*dance instruction manuals, anti-dance manuals, histories, treatises on etiquette" in *Description* vs."Ballroom dancing—United States" in *Subjects* (*Subjects* information specifies *Description* information from "dance" to "ballroom dancing" and adds geographic coverage information, while *Description* field adds information on specific aspects of dancing — "etiquette" — and genre of materials in collection not covered by any other metadata field in this record).
- "towns of Coal City, Braidwood, and Wilmington" in *Description* vs."Illinois (state), Grundy (county)" in *Geographic Coverage* (state and county information in *Geographic Coverage* and town information in *Description* complement each other for a more specific geographic representation).
- "Contemporary", "European age of chivalry", "prior to 1900" in *Description* vs. "1200-1900" in *Temporal Coverage* (while *Temporal Coverage* specifies the beginning of the "prior to 1900" range of years — "1200" — and provides the time frame for "European age of chivalry," *Description* introduces another — "contemporary" — time period not covered by *Temporal Coverage*).
- "newspaper photographs" in *Description* vs. "photographs/slides/negatives, archival finding aids" in *Objects* (*Description* specifies genre information in *Objects* from general "photographs" to "newspaper photographs:, while *Objects* adds another genre not mentioned in *Description* — "archival finding aids").

4.2.5 Redundancy between Collection Metadata Fields

While exhibiting the most mutual (two-way) complementarity between the collection metadata fields, The European Library was also the only aggregation that had a noticeable proportion (19%) of redundancy in the collection records in the sample. Very little redundancy was observed in the Opening History and American Memory collection records. Examples of this repetition of the same information in different metadata fields include geographic information (e.g., "Netherlands", "Estonia", "Ljubljana" in both *Description* and *Geographic Coverage* fields), and temporal information (e.g., "1763" in both *Description* and *Temporal Coverage* fields), as well as genre information (e.g., "photographs" in both *Description* and *Subjects*).

4.3 Summary and Discussion of Collection Metadata Findings in Relation to Existing Best Practice Guidelines

As results of the content analysis show, collection-level metadata in the three aggregations exhibits high levels of metadata richness. The collection metadata richness includes such components as

- representation of collection's subject matter with mutually-complementary values in different metadata fields and
- variety of collection properties/characteristics encoded in the free-text *Description* field.

A total of 19 different collection characteristics were found in free-text *Description* fields across the three aggregations. The free-text *Description* collection metadata field was found to contain on average 6 different collection properties or characteristics. Types and genres of

objects in a digital collection, topical subjects, geographic and temporal coverage were found to be the most consistently represented collection characteristics. Additional five collection characteristics were found only in the free-text *Description* fields, and in no other collection metadata field: the creator of items in a digital collection, the provenance, the uniqueness, importance, and comprehensiveness of content in a digital collection.

The information found in different collection metadata fields is often mutually complementary. The assumption, based on the pilot studies, that the *Description* field would often complement other metadata fields, was supported by this study's findings. It was also observed in this study that information in other collection metadata fields complements information in *Description* field almost as often, sometimes even more often (as in the case with the *Geographic Coverage* field).

This study observed certain differences in the application of collection metadata fields and distribution of collection properties in free-text *Description* fields, which might be explained by the specifics of the policies (including collection development policies) followed and tools used in describing digital collections in the three aggregations. Opening History was found to be the most consistent in applying all four of the subject-specific collection metadata fields — *Subjects*, *Temporal Coverage*, *Geographic Coverage,* and *Objects*, while the collection records in the other two aggregations often lacked *Temporal Coverage* or *Geographic Coverage* fields. American Memory was found to have both longer and richer free-text *Description* fields overall, with 14 collection properties encoded in this field more consistently than in Opening History and European Library. The European Library was found to have the shortest free-text *Description* fields overall but to encode provenance, and frequency of additions information more consistently that the other two aggregation. Geographic coverage and creator of items

information were found more often in free-text *Description* fields in Opening History, while information about intended audience or uses of a digital collection was found equally more often in Opening History and The European Library than in American Memory. More mutual (two-way) complementarity but also more redundancy between the values in different collection metadata fields was observed in European Library. Despite the differences observed in three large-scale aggregations, most of their free-text *Description* collection metadata fields were found to provide rich description of digital collections, covering a variety of collection properties and complementing information encoded in other collection metadata fields.

The findings presented in this chapter demonstrate the richness of collection metadata records in large-scale aggregations of digital collections. Results of this study indicate that encoding of mutually complementary subject-specific information in free-text and controlled vocabulary metadata fields is already being recognized as a benchmark in crafting rich collection-level metadata in aggregations. In addition to subject-specific information, the emerging best practices in collection-level description observed in this study suggest enriching *Description* fields by encoding a variety of other collection characteristics such as title, size, collection development policy, copyright information, provenance, intended audience, navigation and functionality, language of items in collection, frequency of additions, participating or contributing institutions, funding sources, and especially the characteristics for which no specialized collection metadata fields exist: collection strengths (importance, uniqueness, and comprehensiveness) and creators of items in collection.

Table 9 compares the findings of the content analysis of free-text *Description* fields with the best practice recommendations for collection-level *Description* fields and applicable item-level guidelines derived from sources including *Cataloging Cultural Objects* (CCO, 2008), *OSU*

Knowledge Bank Metadata Application Profile (OSU, 2006), *National Union Catalog of Manuscript Collections* (NUCMC), and *Online Audiovisual Catalogers Cataloging Policy* guidelines (OLAC, 2002). This comparison makes it clear that while meeting most of the recommendations, collection-level *Description* fields in Opening History, American Memory, and The European Library aggregations also routinely include 7 additional kinds of information about digital collections that are not covered by these recommendations: comprehensiveness, copyright, frequency of additions, funding sources, hosting/contributing institution, size, and title. Encoding these additional collection properties in *Description* fields might be considered an emerging best practice that is not yet reflected in any of the guidelines documents. Several recommendations, which might be very specific and apply only to a small proportion of digital collections, — information about "consequence, products" (OLAC), "work's relationship to other works, any aspects of work that might be either disputed or uncertain" (CCO), and "particular items of extraordinary interest" (NUCMC) — were not found to be implemented in the sample of *Description* fields in three aggregations.

4.4 Figures and Tables

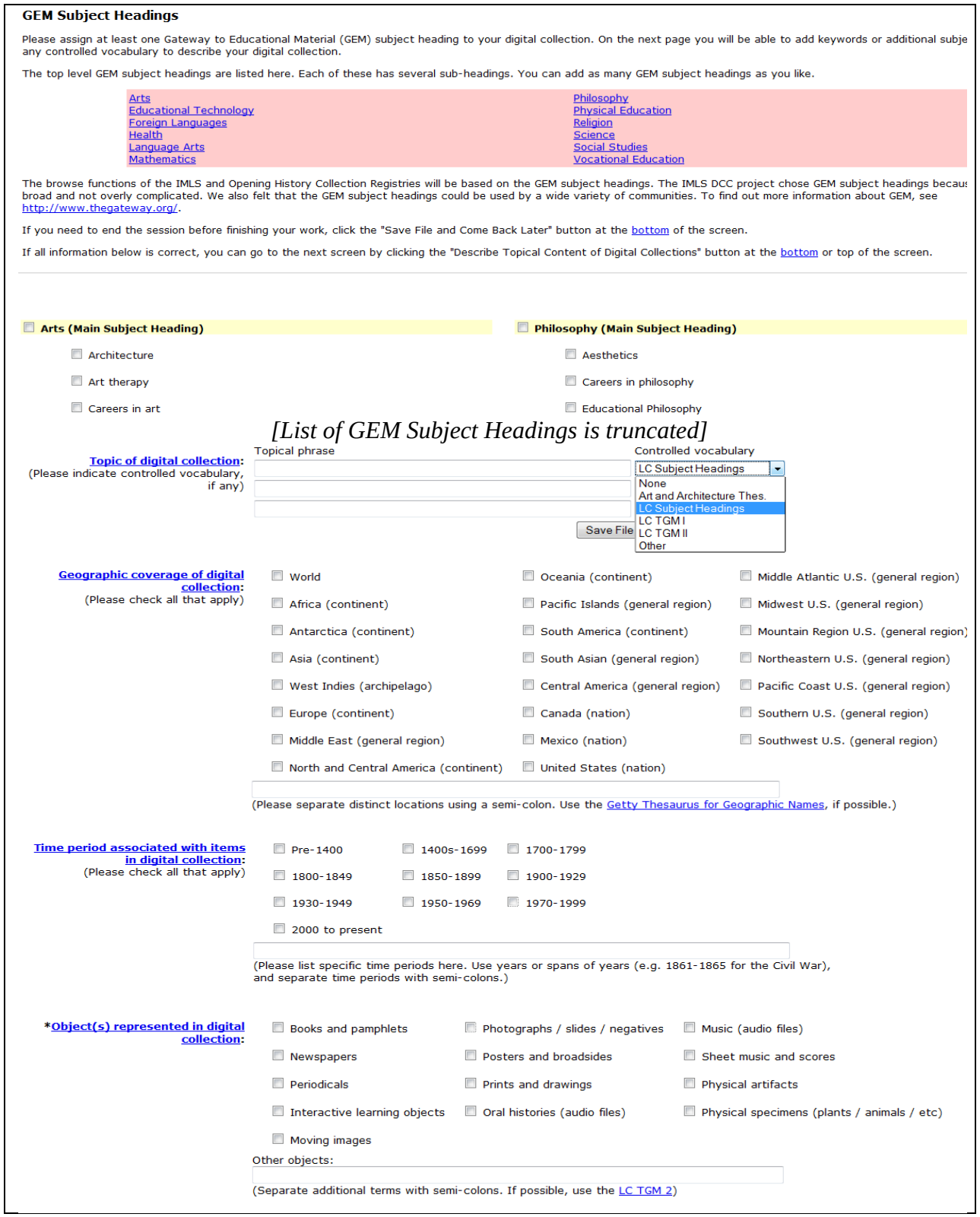


Figure 5. Subject-specific collection metadata fields in Opening History Collection Registry Entry Form

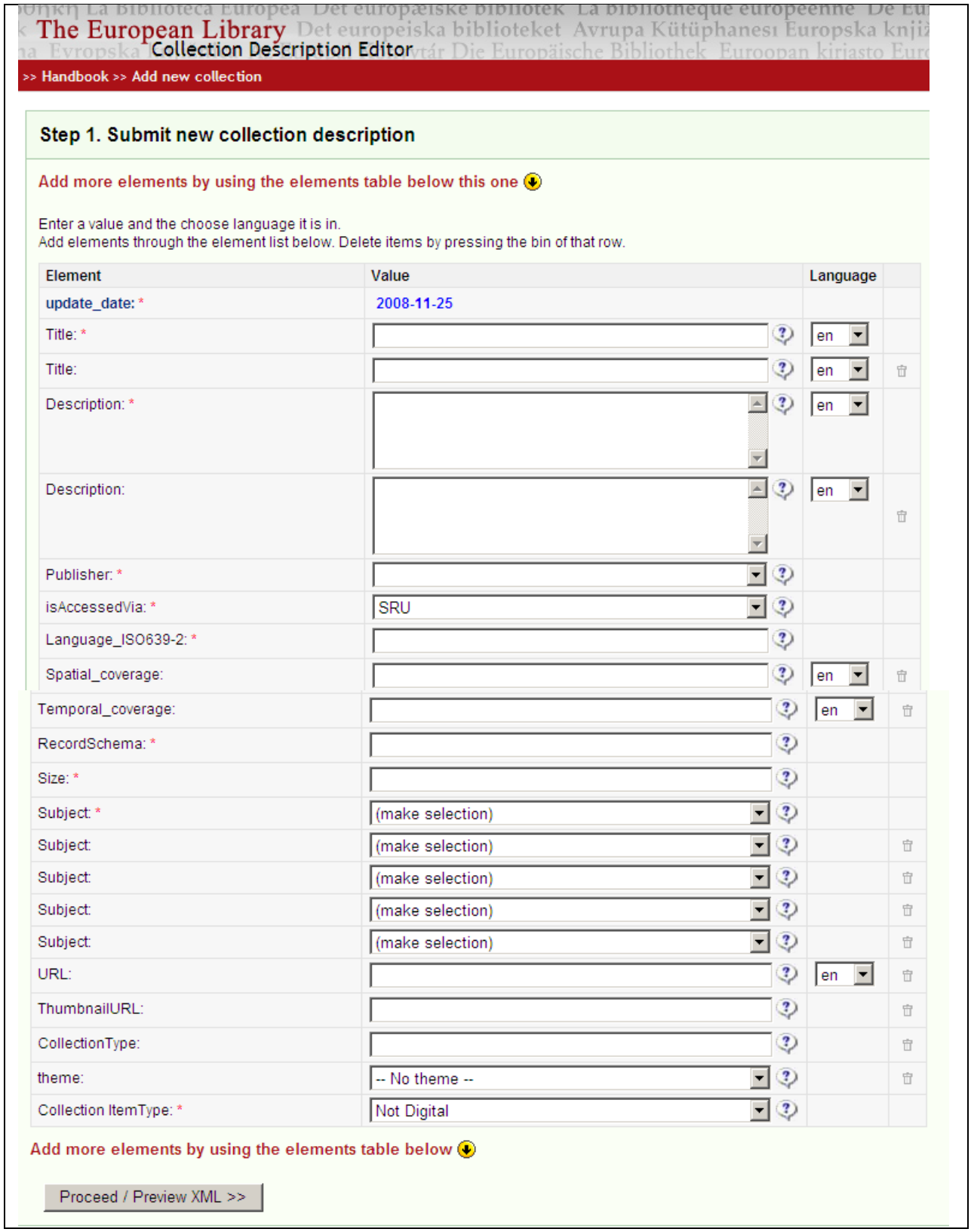


Figure 6. Subject-specific metadata fields in The European Library Collection Description Editor

```
- <subject authority="lcsh">
  - <name type="corporate">
      <namePart>United States.</namePart>
      <namePart>Army.</namePart>
      <namePart>Pennsylvania Infantry Regiment, 105th (1861-1865)</namePart>
    </name>
  </subject>
- <subject authority="lcsh">
    <topic>Autograph albums</topic>
  </subject>
- <subject authority="lcsh">
    <geographic>United States</geographic>
    <topic>History</topic>
    <temporal>Civil War, 1861-1865</temporal>
    <topic>Campaigns</topic>
  </subject>
- <subject authority="lcsh">
    <geographic>United States</geographic>
    <topic>History</topic>
    <temporal>Civil War, 1861-1865</temporal>
    <topic>Personal narratives</topic>
  </subject>
- <subject authority="lcsh">
    <geographic>United States</geographic>
    <topic>History</topic>
    <temporal>Civil War, 1861-1865</temporal>
    <topic>Psychological aspects</topic>
  </subject>
- <subject authority="lcsh">
    <geographic>Virginia</geographic>
    <topic>History</topic>
    <temporal>Civil War, 1861-1865</temporal>
    <topic>Campaigns</topic>
  </subject>
- <subject authority="itoamc">
    <occupation>Soldiers</occupation>
```

Figure 7. Granularity of subject-specific collection metadata: American Memory

```
<dc:coverage>Southern U.S. (general region)</dc:coverage>
<dc:coverage>North and Central America (continent)</dc:coverage>
<dc:coverage>United States (nation)</dc:coverage>
<dc:coverage>Georgia (state)</dc:coverage>
<dc:coverage>Athens, Georgia (city)</dc:coverage>
<dc:coverage>Clarke County, Georgia (county)</dc:coverage>
<dc:coverage>1900-1929</dc:coverage>
<dc:coverage>1906</dc:coverage>
<dc:subject>Social Studies--Human relations</dc:subject>
<dc:subject>Social Studies--State history</dc:subject>
<dc:subject>Social Studies</dc:subject>
<dc:subject>University of Georgia--Directories</dc:subject>
<dc:subject>University of Georgia--Faculty--Directories</dc:subject>
<dc:subject>University of Georgia--Alumni and alumnae--Directories</dc:subject>
<dc:subject>University of Georgia--History</dc:subject>
```

Figure 8. Granularity of subject-specific collection metadata: Opening History

```
<dcterms:spatial xml:lang="en">Netherlands</dcterms:spatial>
<dcterms:spatial xml:lang="en">Indonesia</dcterms:spatial>
<dcterms:temporal xsi:type="dcterms:period" xml:lang="en">19th century, 20th
  century</dcterms:temporal>
<dc:subject xml:lang="en">900 History and geography</dc:subject>
<dc:subject xml:lang="en">300 Social sciences</dc:subject>
<dc:subject xml:lang="en">779 Photographs</dc:subject>
```

Figure 9. Granularity of subject-specific collection metadata: European Library

```
<dc:title xml:lang="en">Indonesia independent - Dutch photographs 1947-
  1953</dc:title>
<dc:description xml:lang="en">The decolonisation of the Dutch Indies in more than
  4500 photographs by Cas Oorthuys, Charles Breijer and Lex de Herder from the
  collection of the KITLV</dc:description>
<dc:title xml:lang="nl">Indonesië onafhankelijk - foto</dc:title>
<dc:description xml:lang="nl">De dekolonisatie van Nederlands-Indië in meer dan
  4.500 foto's van Cas Oorthuys, Charles Breijer en Lex de Herder uit de collectie
  van het KITLV.</dc:description>
<dc:title xml:lang="bg">Индонезия независима - холандски фотографии 1947-
  1953</dc:title>
<dc:description xml:lang="bg">Деколонизацията на Холандските Индии в повече от
  4500 фотографии от Cas Oorthuys, Charles Breijer и Lex de Herder от колекцията
  на KITLV</dc:description>
<dc:title xml:lang="cs">Indonésie nezávislá - nizozemské fotografie 1947-
  1953</dc:title>
<dc:description xml:lang="cs">Dekolonizace Nizozemské východní Indie na více než 4
  500 fotografií, jejichž autory jsou Cas Oorthuys, Charles Breijer a Lex de Herder,
  ze sbírky KITLV</dc:description>
<dc:title xml:lang="da">Indonesien selvstændig - hollandske fotografier 1947-
  1953</dc:title>
<dc:description xml:lang="da">Afkolonialiseringen af Hollandsk Østindien i mere end
  4500 fotografier af Cas Oorthuys, Charles Breijer og Lex de Herder fra samlingen
  på KITLV</dc:description>
<dc:title xml:lang="de">Unabhängiges Indonesien - Niederländische Fotografien von
  1947-1953</dc:title>
<dc:description xml:lang="de">Die Entkolonialisierung der niederländischen Kolonien in
  Ostindien wird in mehr als 4.500 Fotografien von Cas Oorthuys, Charles Breijer und
  Lex de Herder in der Sammlung des KITLV festgehalten.</dc:description>
```

Figure 10. Multilingual collection metadata in The European Library

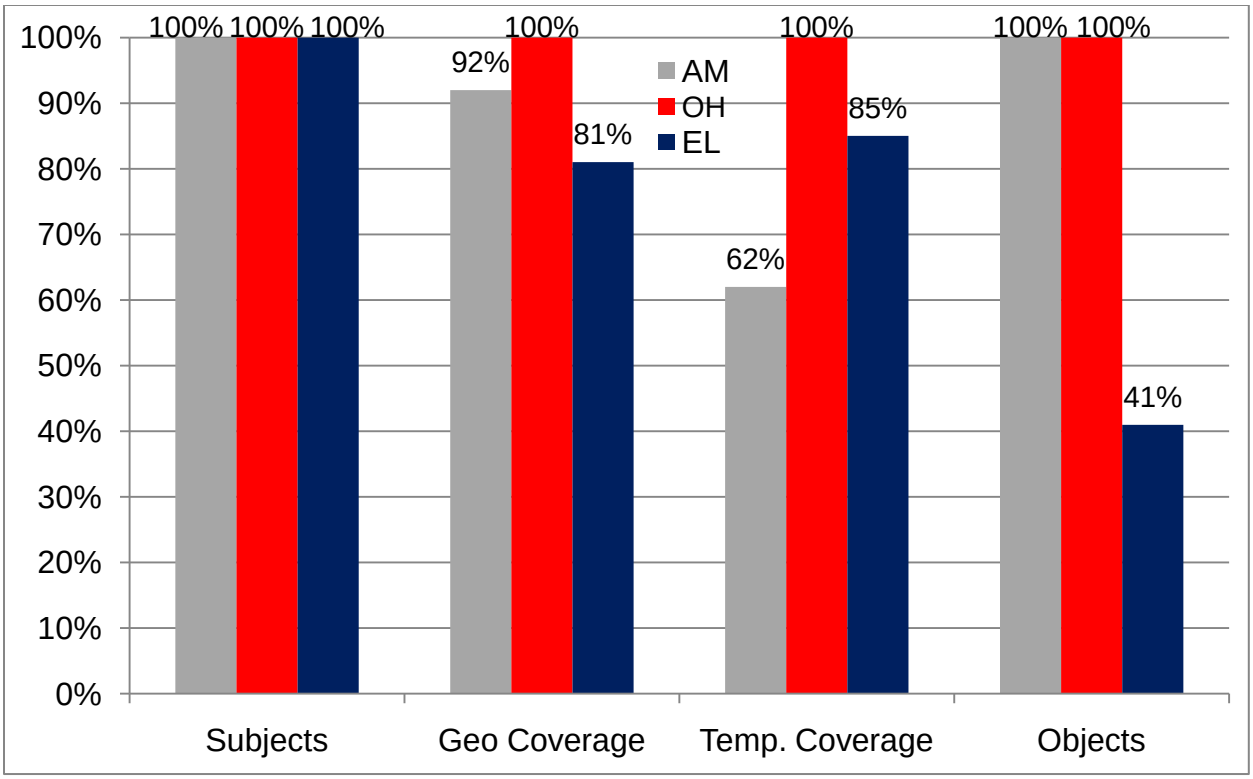


Figure 11. Application of subject-specific collection metadata fields in three aggregations

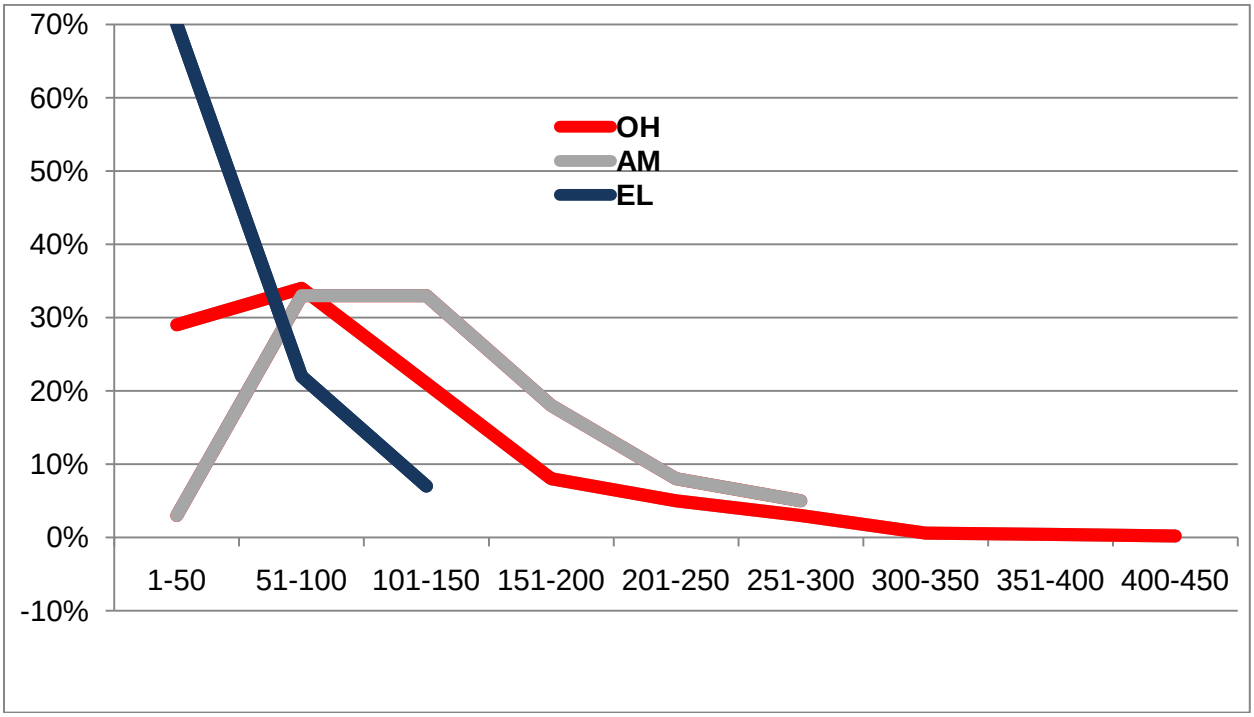


Figure 12. Distribution of *Description* field length in Opening History, American Memory, and The European Library aggregations

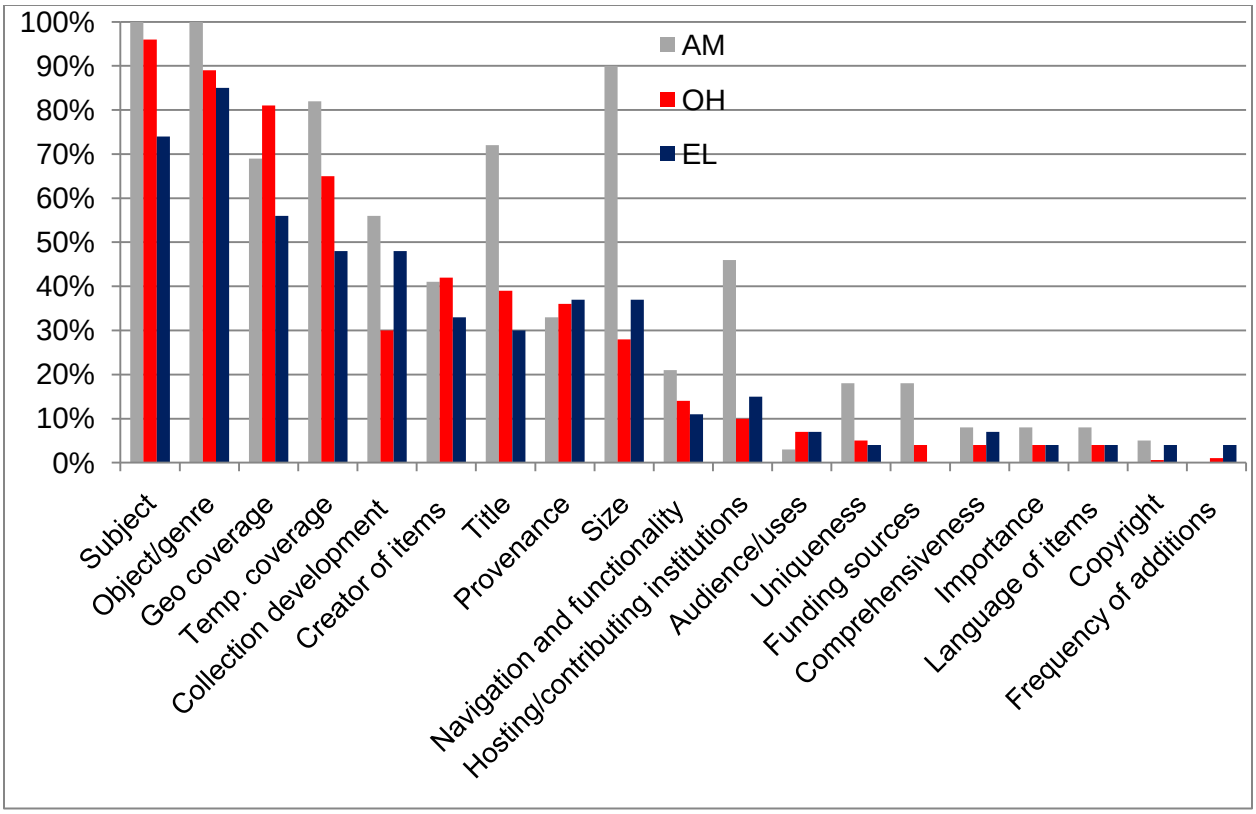


Figure 13. Distribution of collection properties in *Description* fields in three aggregations

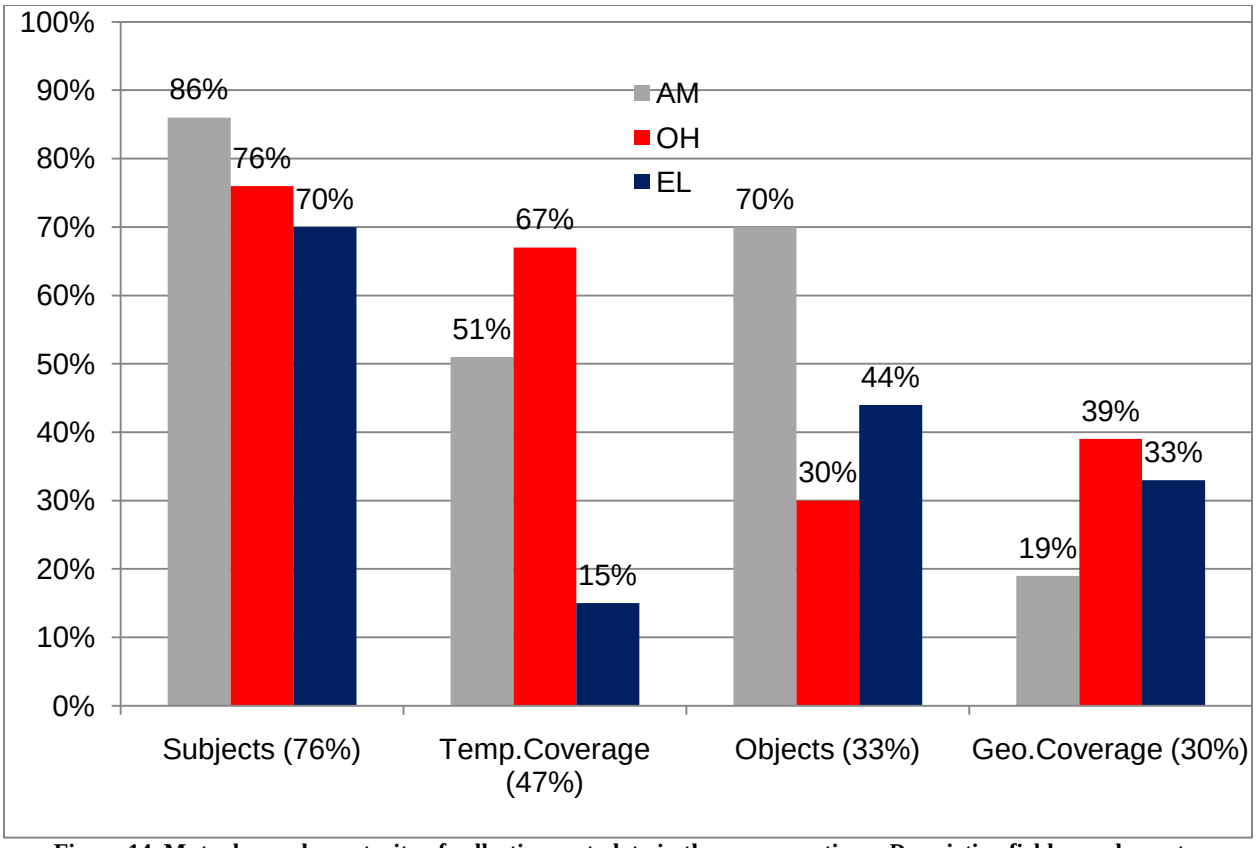


Figure 14. Mutual complementarity of collection metadata in three aggregations: *Description* field complements information in other subject-specific fields

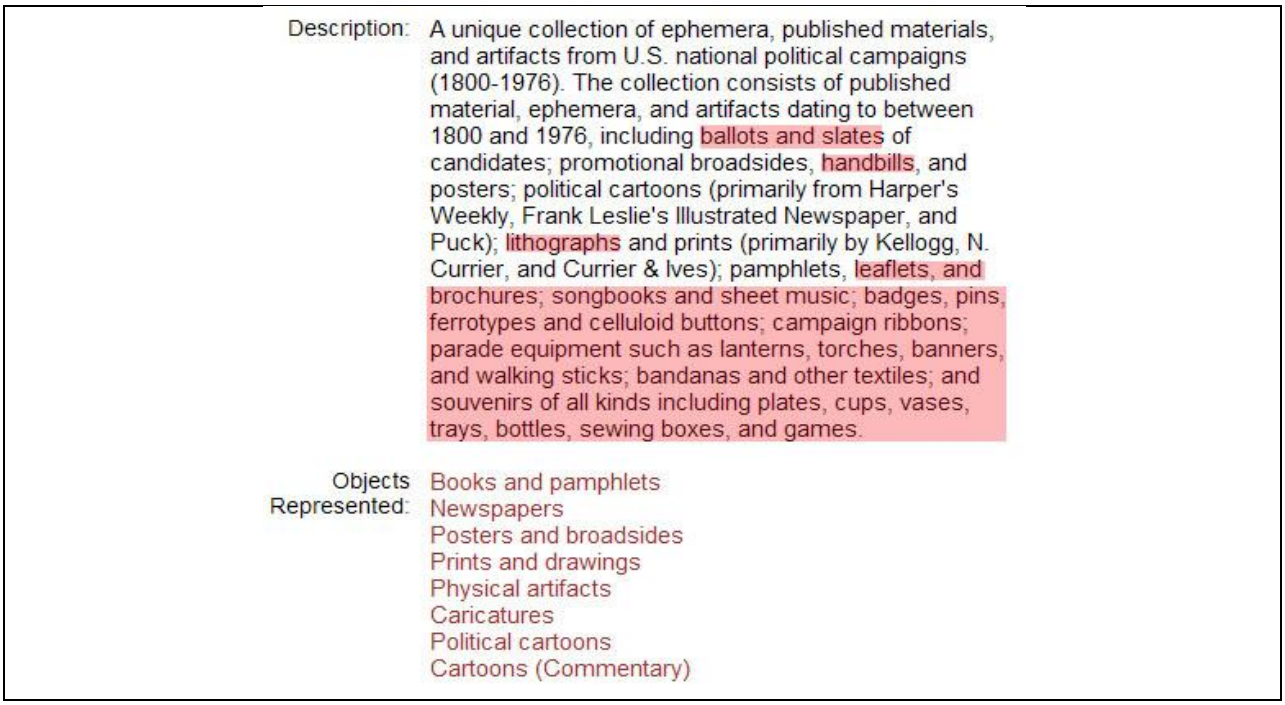


Figure 15. Object types information in *Description* field

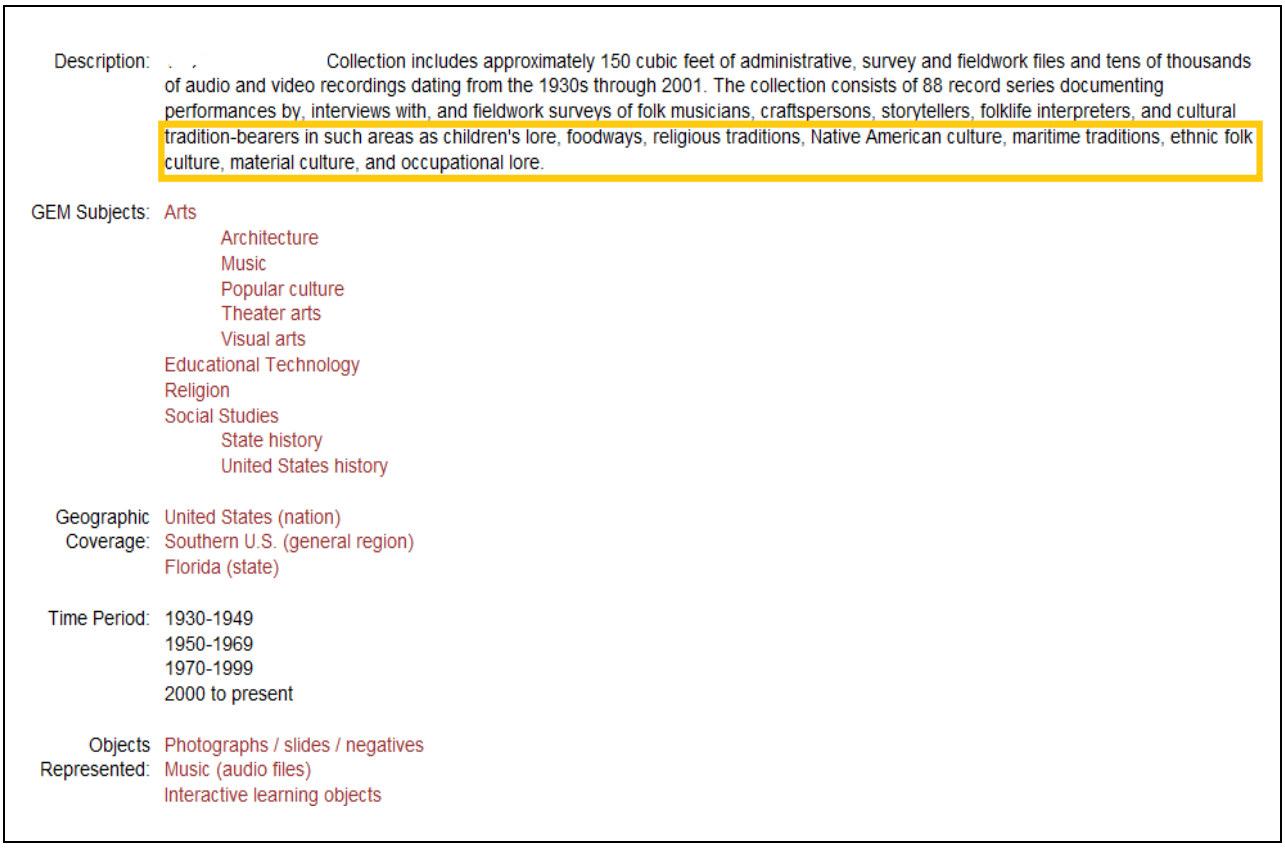


Figure 16. *Description* field complementing multiple subject-specific fields

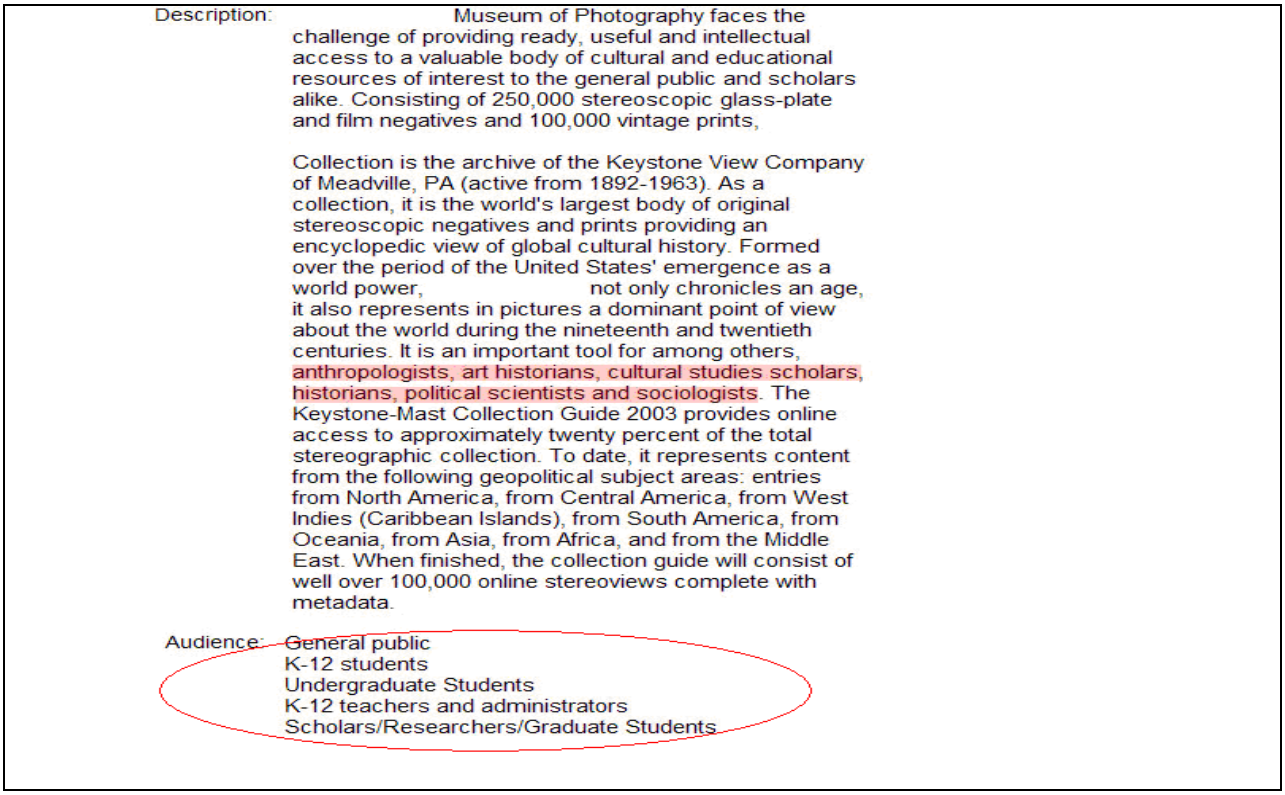


Figure 17. Audience information in *Description* field

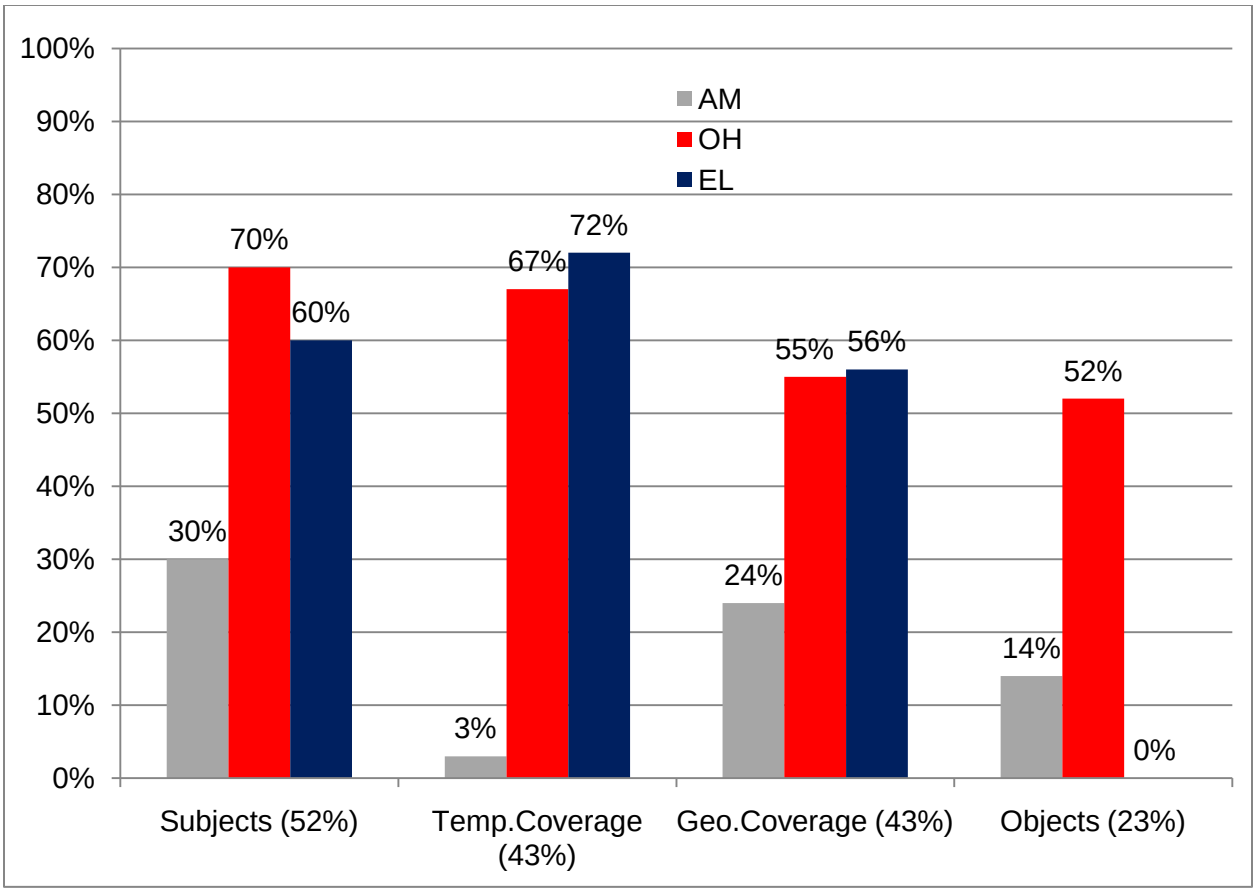


Figure 18. Mutual complementarity of collection metadata in three aggregations: other subject-specific fields complement information in *Description* field

	OH (DCCAP)	**AM** (MODS)	**EL** (ELAPCD)
Subjects	<dc:subject> -GEM Subjects -[Other] Subjects -Library of Congress Subject Headings	<topic> <name> <occupation>	<dc:subject>
Objects	<dc:type> Objects Represented	<type of resource> <genre>	<theme>
Geographic Coverage	<dc:coverage> Geographic Coverage	<geographic>	<dcterms:spatial>
Temporal Coverage	<dc:coverage> Time Period	<temporal>	<dcterms:temporal>

Table 3. Mapping subject-specific collection metadata fields in American Memory, Opening History, and The European Library

Aggregation	Mean	Median	Variance	Standard deviation
American Memory (n=39)	132	120	3105	55.73
Opening History (n=488)	98.2	82.5	4861	69.72
The European Library(n=27)	45.1	43	725.2	26.23

Table 4. *Description* field lengths in three aggregations: variability measures

Field	Aggregation	Mean	Median	Variance	Standard deviation
Subjects	American Memory (n=37)	21.4	8	1533.1	39.15
	Opening History (n=33)	19.94	11	459.5	21.44
	The European Library (n=27)	8.81	7	52.62	7.25
Temporal Coverage	American Memory (n=37)	10.05	3	448.52	21.18
	Opening History (n=33)	6.88	6	18.61	4.31
	The European Library (n=27)	6.17	12.5	60.06	7.75
Geographic Coverage	American Memory (n=37)	3.29	2	5.06	2.25
	Opening History (n=33)	13.82	15	38.15	6.18
	The European Library (n=27)	1.41	1	0.54	0.73
Objects	American Memory (n=37)	6.11	2	160.65	12.67
	Opening History (n=33)	5.12	3	16.297	4.04
	The European Library (n=27)	1	1	0	0

Table 5. Lengths of subject-specific collection metadata fields in three aggregations

Aggregation	Mean	Median	Variance	Standard deviation
American Memory (n=37)	171.75	144	8158.25	90.32303
Opening History (n=33)	144	131	6846.44	82.74
The European Library (n=27)	61.15	50	799.52	28.28

Table 6. Cumulative lengths of *Description* and subject-specific collection metadata fields (*Subjects*, *Temporal Coverage*, *Geographic Coverage*, *Objects)* in three aggregations

Aggregation	Mean	Median	Variance	Standard deviation
American Memory (n=39)	7.77	8	3.39	1.84
Opening History (n=488)	5.62	6	3.09	1.76
The European Library (n=27)	5.04	5	2.34	1.53

Table 7. Number of collection properties encoded in free-text *Description* collection metadata fields in three aggregations

Collection property	American Memory	Opening History	European Library	Cumulative %
Subject	100%	96%	74%	95%
Object/genre	100%	89%	85%	90%
Geographic coverage	69%	81%	56%	79%
Temporal coverage	82%	65%	48%	65%
Creator of items	41%	42%	33%	42%
Collection development	56%	30%	48%	42%
Title	72%	39%	30%	41%
Provenance	33%	36%	37%	36%
Size	90%	28%	37%	33%
Navigation and functionality	21%	14%	11%	15%
Hosting, contributing, participating institutions	46%	10%	15%	13%
Audience/uses	3%	7%	7%	7%
Uniqueness	18%	5%	4%	6%
Funding sources	18%	4%	0%	5%
Comprehensiveness	8%	4%	7%	5%
Importance	8%	4%	4%	4%
Language of items	8%	4%	4%	4%
Frequency of additions	0%	1%	4%	1%
Copyright	5%	1%	4%	1%

Table 8. Distribution of collection properties in *Description* fields in three aggregations: more details

	Existing Guidelines on Information to Include in *Description* Field				This study
	OLAC Cataloging Policy	***Cataloging Cultural Objects***	***OSU Knowledge Bank Metadata Application Profile***	***National Union Catalog of Manuscript Collections***	**collection properties found in *Description* fields**
METADATA GUIDELINES	specific types and forms of materials present	N/A	N/A	types of materials included in the collection	Object types/genres
	significant people and topics covered		N/A	topics with which the materials in the collection deal	Subjects
	significant places covered	N/A	N/A	geographical areas, with which the materials in the collection deal	Geographic coverage
	significant events covered, span of dates covered by the collection	N/A	N/A	associated dates, events, and historical periods dealt with by the materials in the collection	Temporal coverage
	history of the work	N/A	provenance, history of the work	N/A	Provenance
	N/A	significance	N/A	N/A	Importance
	unique characteristics of the collection	N/A	N/A	N/A	Uniqueness
	reason and function of the collection	N/A	N/A	N/A	Collection development policy
	N/A	N/A	N/A	names, dates, and biographical identification of persons and names of corporate bodies significant (by quality and/or quantity of material) to the collection, specific phases of career/activity of the major person/body responsible	Creator of items in the collection (corporate or individual)

Table 9. Best practices in collection-level *Description* field: existing guidelines and findings of this study

Table 9 (cont.)

	Existing Guidelines on Information to Include in *Description* Field				**This study**
MET GUIDELINES	***OLAC Cataloging Policy***	***Cataloging Cultural Objects***	***OSU Knowledge Bank Metadata Application Profile***	***National Union Catalog of Manuscript Collections***	**collection properties found in *Description* fields**
	N/A	N/A	nature of the language of the resource	N/A	Language of items
	audience	N/A	N/A	N/A	Audience/uses
	user interaction	N/A	N/A	N/A	Navigation and functionality
UNMET GUIDELINES	consequence, products	1. relationship to other works [or collections] 2. any aspects of work[or collection] that might be either disputed or uncertain		particular items of extraordinary interest	N/A
EMERGING PRACTICES	N/A	N/A	N/A	N/A	**1.** Comprehensiveness **2.** Copyright **3.** Frequency of additions **4.** Funding sources **5.** Hosting/contributing institution **6.** Size **7.** Title

Chapter 5. User Interactions with Aggregations of Digital Collections: Findings and Discussion

This chapter presents results of investigation into user interactions with Opening History aggregation of digital collections, which were obtained by applying two research methods: transaction log analysis, and interview and observation sessions with scholars interacting with Opening History and American Memory aggregations. The findings presented in this chapter answer the following research questions:

- ***How do scholarly users of cultural heritage aggregations approach collection-level information discovery?***
- ***Which collection-level metadata fields provide scholarly users with the valuable information to meet their needs?***
- ***How does collection-level user search data fit the FRBR model?***

5.1 Transaction Log Analysis Findings

This section presents results of transaction log analysis of user interactions with Opening History, supplemented with the data on user interactions with its predecessor — the IMLS Digital Collections and Content aggregation — over the period of one year, from February 2008 to January 2009. Please consult Chapter 3 for detailed discussion on the data sampling, collection, processing, and analysis.

5.1.1 Major Types of User Interactions with Opening History

The transaction log analysis shows that browsing — both collection- and item-level — is used more often than search. Collection-level user interactions — search and browse —occurred

overall almost as often as item-level interactions. Only 7% of user interactions are collection searches. Table 10 and Figure 19 provide basic page-view statistics for browse and search interactions.

5.1.2 Collection Browse in Opening History

The analysis indicates (Figure 20) that subject browse is used most often (32%) among the faceted browse options provided in the Opening History and/or IMLS DCC aggregations. Browse by geographic area and type of objects in collection also proved popular (18% and 17% of collection browse respectively). Users also browsed by project (17%) and institution (11%) responsible for the digital collection, and by collection title (5%).

Statistical characteristics of query frequencies for various types of collection browse queries are summarized in the Table 11. The median frequency for the browse query equals 1 for most browse types. Median project browse frequency is somewhat higher — 2. Although collection title browse was recorded the least often among collection level browse types, both mean and median query frequency is the highest in this query type. The highest standard deviation of query frequency was also exhibited by the title browse.

5.1.3 Pageviews of Collection and Item Metadata Records

As shown in Table 12, in Opening History the collection-level view of metadata records was performed 1,760 times, more often than any other collection-level user interaction, and almost as often as item search. A total of 235 collection records were viewed from 1 to 42 times each. This means that most of the collection-level queries were followed by viewing at least one collection-level metadata record. Average query frequency for collection metadata views was 3.1, while the median query frequency was 2, with a variance of 7.1 and a standard deviation of

2.67. By contrast, item metadata records were viewed almost 80% less than collection metadata records — only 368 times.

This indirect indication of the value of collection metadata to the users of the cultural heritage aggregation is supported by the findings of interviews and observations with the United States history scholars, which show that historians consider descriptions of collections as a whole important and useful for their exploration of digital content in aggregations of digital collections. Participants of this study almost always view collection records in the process (see Section 5.2 for detailed discussion on this). Although these scholars usually know which collections can be found in more familiar digital resources (e.g., American Memory) and often skip viewing collection metadata records in this highly-familiar environment, they examine the collection records in less familiar environments or rapidly growing resources (e.g., Opening History), for which they have not yet formed clear collection expectations. Collection metadata records also sometimes help scholars learn that a familiar, previously used physical collection, or a part of that collection, is now digitized and available online.

5.1.4 Collection and Item Searches in Opening History

The analysis of collection search and how it compares to item search was the focus of the transaction log analysis of user interactions with an aggregation of digital collections in this study. Quantitative and qualitative analysis of various patterns of searching behavior in aggregation included the analysis of search frequency, search approaches, query length, and query frequency.

5.1.4.1 Search Approaches and Search Terms

The analysis of the web log data shows that the collection search was used overall less often than item search. Over the 12 week period in the sample, collection search was initiated 880 times (Figure 21). Collection search frequency varied throughout the year. April had the lowest collection search activity (4.28% of collection searches). May and October were found to be the two busiest months in collection search, with 24.72%, and 14.31% of all collection searches in a sample respectively. The only advanced collection-level search option provided by the Opening History aggregation, which allows users to limit collection search results by the type of objects in digital collections (for example, to retrieve only collections that contain oral histories), was used 79 times (8.97% of collection search queries), the remaining 91.3% of collection-level searches used the basic keyword search option. The phrase collection search was attempted 14 times (1.59 % of collection search queries). No attempts to use Boolean operators to build the collection search query were observed.

The item search was used more than twice as often as collection search. Over the 12 week period in the sample the item search was conducted 1,860 times. Most of the item searches in the sample were basic keyword searches, while advanced search option was used in 574 cases (30.8 % of all item search instances). Advanced item search was used in Opening History much more often than is usually observed in the Web search studies using transaction log analysis (e.g., Spink & Jansen, 2004). Interestingly, the recent study of the use of the similar The European Library aggregation (Agosti et al., 2008), which used questionnaire as its research method, reported that "81% of users prefer the advanced search facilities" (p. 42) — significantly higher numbers than the body of search data in this transaction log analysis study of Opening History suggests. A total of 243 user queries (42.3% of advanced search instances or

13% of all item search instances) in the log used the fielded search — searched by author (92, or 37.8% of fielded searches), title/subject words (148, or 60.9% of fielded searches), or combined author and title/subject words fields (3, or 1.2 % of fielded searches). The option to limit item search to a specific collection or a group of collections was used in 138 cases (24 % of advanced search instances or 7.4% of all item search instances). The phrase item search was attempted 11 times. No attempts to use Boolean operators to build the item search query were observed. Slightly over 5% of item search queries were formulated in languages other than English: mostly in French (e.g., “Les Mannequins Politiques. Ce jeu n'a duré que trois jours”) and German (e.g., “er hat die wandlung gebracht”), but also in Italian (e.g., “arditi”), Latin (“emblemata morale”), and Dutch (e.g., “uitspraak 2005 raad van state”).

5.1.4.2 Search Query Length and Frequency

Figure 22 displays distributions of search query length at the collection and item level. As can be seen from the Table 13, the collection search query length ranged from 1 to 7 words per query. The average collection search query length in the Opening History aggregation was found to be shorter than the average item query length — 1.75 words per query — while the median was recorded at only 1 word per query. The length of item search queries ranged from 1 to 15 words per query. The average item search query length was found to be 1.99 words per query, while the median was recorded at 2 words per query.

Similarly, the average frequency of the item search query was found to be higher than the frequency of collection search query in the Opening History aggregation. On average, the item search query was used 1.99 times, while the collection search query was used only 1.54 times. In both item-level and collection-level searches, a large majority of queries occurred only once: 78% of item queries and 68% of collection queries. Item search query frequency exhibited a

much higher variability (variance of 12.99 and standard deviation of 3.60) than collection search query (variance of 1.31 and standard deviation of 1.14).

5.1.4.3 Search Categories

One of the aims of this research was discovering how the Functional Requirements for Bibliographic Records (FRBR) entity-relationship model of bibliographic universe fits collection-level subject searching by scholarly historians. The more specific research questions addressed included:

- What is the distribution of search categories in collection-level user searching in Opening History aggregation?
- What (if any) are the categories not covered by FRBR model and previous analysis?
- How do the distributions of the FRBR-based search categories compare at collection-level and item-level?

FRBR set of 10 entities served as a basis or analytical framework for search categorization in this study. Unique search queries identified in the web log data were categorized using the FRBR-based list of search categories. Coding of the user keyword searches was based on the procedure described by the Coding Manual (Appendix C).

Weak-to-medium positive correlation was observed between the length of the search query and the number of the FRBR-based search categories to which it belongs (Pearson R of 0.3287 for item search queries and 0.40609 for collection search queries). Collection and item search queries exhibited very similar specificity: on average, a collection search query covered 1.3992 search categories, while an item search query covered 1.3955 FRBR-based categories.

The median number of search categories per search query was the same in both cases — 1. Slightly over a third of search queries — 35.5 % of collection search queries and 34.6 % of item search queries — belonged to 2 or more categories, for example, "Watkins cookbooks" (*person* and *object*), "early 1800 homes Danville" (*event*, *object*, and *place*).

As shown in Figure 23, the top three search categories at the collection level were *object* (e.g., "drinking vessel") with 36% of searches, *place* (e.g., "Chile") with 26% of searches, and *concept* (e.g., "civil right") with 22% of searches. It is worth noting that these three categories belong to FRBR Group 3 of or subject entities. However, the fourth FRBR Group 3 subject entity — *event* (e.g., "1935 meat strike") — was observed in collection searches much less than the other three (9%). The FRBR Group 2 search categories — *person* (e.g., "Alfred R. Glancy Jr.") and *corporate body* (e.g. "Dana College", "Kapa Alpha Psi") — were observed somewhat less often than *object*, *place*, and *concept*, in 19% and 13% respectively of the collection searches. The *work/collection* (e.g., "Find It Illinois", "how a colored woman aided john brown") search category was observed in 8% of the searches, while the *ethnic group* (e.g., "Cheyenne") and *class of persons* (e.g., "fashion designers") search categories were observed in 3% of collection searches each. No *family* search instances were observed in the sample of web log data.

In the item-level searches (Figure 23), the distribution of search categories was somewhat different: while *object* was found to be the most often used search category (28%), the second most used search category (27%) was *person*, which belongs to FRBR Group 2 of entities, and the third was *place* (24%). Three more search categories were observed more often at the item, than at the collection level: FRBR Group 3 *event* (10% of item searches and 9% of collection

searches), *ethnic group* (4% of item searches and 3% of collection searches) and *class of persons* (7% of item searches and 3% of collection searches).

A total of 45 unique search queries were performed both in the collection and item searches. As the Figure 24 shows, *object*, *concept*, and *place* are also the top three search categories among these overlapping search queries (i.e., the queries found both in item search log data and collection search log data). However, in this group of queries, *concept* search (38% of queries) is much more prominent than in general at either collection or item level (22% and 17% respectively). No explanation for this finding was provided by the interview and observation data.

Figure 25 displays the distribution of successful (i.e., retrieving at least one collection record in the Opening History aggregation) collection-level user search terms by the corresponding FRBR-based search category. This chart indicates that *object, concept*, and *place* searches in the sample were the most often used among successful collection-level search queries. These are the same three search categories that were found to be the most often occurring among all the searches (Figure 23) and among the search queries used both in collection and item searches (Figure 24). The next section reports more results on the successful collection-level searches in Opening History, with a focus on the collection metadata fields which contained the matches to user search terms.

5.1.4.4 Collection Metadata Matching User Search Terms

One of the goals of this study was to determine which collection metadata fields provide the most matches to the collection-level user search terms in aggregations, as part of answering the research question:

Which collection-level metadata fields provide scholarly users with the most valuable information to meet their needs?

When the collection-level user searches documented by the web logs were repeated in the Opening History aggregation in Fall 2009, a total of 186 (or 38.75%) unique collection-level queries successfully retrieved at least one collection record. Depending on how broad the search terms were, these searches retrieved between 1 and 736 collection records. Searches for the broader terms like "United States", "Southern", "Social", "government", "war", "documents", "1930," and "1800" each retrieved 100 or more collection records, while a total of 56 more specific search terms like "Burroughs Adding Machine," "Detroit News photo", "League of Nations" etc. retrieved only one collection record each. The average number of retrieved collection records was 22.49, while the median was only 4. In 61.4% of these searches at least one collection record retrieved contained matches in multiple collection metadata fields. The number of collection metadata fields with a match to user search terms in the search results set ranged from 1 to 5, with the average of 1.54 matching fields per collection record.

Table 14 lists the collection metadata fields in descending order by the relative number of matches to user search terms. *Description* and *Subject* provided the most matches to the user search terms and seemed to satisfy most of the user search queries. *Objects* and *Geographic Coverage* are ranked the sixth and seventh on this list of 23 collection metadata fields describing digital collection, while *Temporal Coverage* is ranked fifteenth.

Matches to the user search terms were also found in five additional areas of collection metadata record which describe related entities such as parent collections and sub-collections, associated physical and digital collections, and the digitization projects associated with collection (Table 15).

As demonstrated by the Figure 26, while the information contained in metadata fields intended for subject and object representation provides a match to a high proportion of user search terms in Opening History, the free-text *Description* field plays the most important role in providing matches to collection-level user searches. In 93% of collection-level searches, at least one of the collection records retrieved would have a match in *Description* field, while 74% of collection searches retrieve one or more collection records with a match exclusively in this field. *Subjects* is another collection-level metadata field that provides a significant source of matches to user search terms, with at least one retrieved collection record having a match to user search term in this field in 50% of searches, and 27% of searches retrieving one or more records with a match exclusively in this field.

Moreover, as demonstrated by the Figure 27, only 7 out of 23 collection metadata fields were found to influence the retrieval of collection records. Twenty-one percent of collection records would not be retrieved in collection searches for the user search terms extracted from the web logs, if *Description* fields were absent in the records, 13% if *Temporal Coverage* were absent, 12% if *Geographic Coverage* were absent, 11% if *Subjects* were absent, 5% if *Objects* were absent, 5% if *Alternative Access* were absent, and 3% if *URL* field were absent. It is important to note that the majority of these metadata fields (5 out of 7) are the fields used for encoding subject-specific information about a digital collection. If only the free-text *Description* field is used in collection metadata records, almost a half (41%) of the collections would not be retrieved in response to subject-specific collection searches in aggregation. This finding indicates great importance of applying a variety of subject-bearing collection metadata fields in describing collections to facilitate subject access in aggregations of digital collections. Even if the fields

other than free-text *Description* are not displayed to the user, the value of the richness of "behind-the-scenes" metadata should not be underestimated.

5.2 Interview and Observation Findings

This section reports the findings of interview and observation sessions conducted with academic historians using two cultural heritage aggregations: Opening History and American Memory. The small sample of U.S. history scholars whose area of research closely correlated with the subject strengths of the two aggregations were interviewed and observed to help answer the following research questions:

- How do scholarly users of cultural heritage aggregations approach collection-level information discovery?
- Which collection-level metadata fields provide scholarly users with the most valuable information to meet their needs?

5.2.1 Familiarity with Aggregations and Digital Collections

The interviews and observations of academic historians demonstrate that historians are very familiar with the American Memory aggregation of digital collections and regularly use it for research and teaching purposes. Historians in the sample have clear collection expectation (cf. Marchionini et al., 1993) in regards to American Memory: they know which digital collections are included and usually do not even spend time looking at collection descriptions for the familiar and previously used digital collections. This is not surprising considering the stability of American Memory, which has been in place for over a decade and is not growing. Opening History, on the other hand, was first launched in the Fall of 2008 and has grown almost five-fold between its inception and the time of the interview sessions. None of the three

participants had prior experience using Opening History or were aware of it before being invited to participate in the study.

Two participants commented on awareness of and familiarity with existing digital collections included in Opening History. One participant closely examined the list of collections with matches on the item search results page and commented that he is familiar with most of these collections but usually accesses them differently; he had not been previously aware of the existence of two collections on this list. Another participant's collection search in Opening History retrieved a collection metadata record for *Indian Peoples of the Northern Great Plains* collection. After seeing "Item records unavailable" note in the collection record (meaning that item metadata from this collection had not been harvested into Opening History) he followed the "collection home" link and noticed that the collection of interest was derived from the Barstow Ledger Drawing physical collection — a collection he knew of and was pleased to discover had been digitized.

5.2.2 Exploring Aggregations

The search observations show that in a familiar aggregation (American Memory), the participants start with a search, while in the new environment (Opening History) they often begin exploration with a browse to familiarize themselves with the content and structure of a resource. One historian in the sample explained that his approach to exploring aggregations or looking for information in other environments (e.g., AP news archive website[54], Densho archive of Japanese American oral histories[55] etc.), varies and often depends on the purpose of exploration: "if I am looking for something specific, I search. A lot of times I browse because I am looking for things

[54] http://www.ap.org/
[55] http://www.densho.org/densho.asp

that might be helpful. I might do just a general search and then just look at everything, or look at what looks interesting and take a look deeper into that collection." In addition, two out of three historians followed links from collection records retrieved in search or browse sessions in Opening History to collection homepages at the institution. In this activity, Opening History is being navigated as a central hub that successfully links the researcher to the originating source. The participants followed leads from the aggregation to explore related digital collections or information about physical collections at the hosting institution.

5.2.2.1 Search in Aggregations

Two participants reported a preference for starting interaction with the aggregations with a search and then browsing the search results. All participants used the basic keyword search option in both Opening History and American Memory. They easily found the prominently located item search window on the Opening History homepage but had difficulty finding the collection-level search window hidden in the bottom of the page. None of the participants used the advanced item search in Opening History during the observation session; only one of them used the quasi-advanced collection search option in Opening History, which allows limiting a search to collections with certain object types, and commented on the usefulness of this option. This observation is consistent with the findings of transaction log analysis performed as part of this research. The transaction log analysis shows that while 30.8% of item searches in Opening History were advanced searches, the option to limit collection-level search results to collections with specific object types was used in only 8.97% of collection searches.

All participants reported using broader search terms for their teaching-related searches (e.g., "Civil War", "slavery") while using more specific terms for their research-related searches:

- relatively narrow topics (e.g., “Japanese American internment”)
- corporate names (e.g., “War Relocation Authority”)
- family names (e.g., “Laublin”, “Pueblo”), or
- tribal names (e.g., “Ogibwe” spelled a couple of different ways, “Paiutes”, “Illinois Indians”, “Crow Indians”).

They also reported that the search terms they use in aggregations or in web browsers like Google are not derived from any controlled vocabulary, however one of the participants commented that he is familiar with Library of Congress Subject Headings in his area and often uses them as search terms in library online catalogs.

5.2.2.2 Browse in Aggregations

Two out of three participants started exploring Opening History with collection-level browse. This may be because the collection browse options are prominently displayed on the homepage of Opening History. One researcher started with the subject browse by selecting “Social studies” in the list of top-level GEM subjects. Another explored two different collection browse options — by place and by subject — before moving to search. After conducting searches, two participants browsed through the search results at both item and collection level before selecting an item or collection record of interest. Yet another researcher explored subject and hosting institution browse options in the middle of the session, after performing item search and viewing collection and item metadata records.

Overall, subject browse was used at least once by each of the three participants during the observation sessions. This observation is consistent with the findings of transaction log

analysis, which show that subject browse was the most often used collection browse option in Opening History, with 32% of all collection browse instances.

Geographic browse option was selected by one observation participant, however this and other participants were also observed clicking on hyperlinked values in *Geographic Coverage* collection metadata fields in Opening History — an action which results in opening a geographic browse result window. At the same, transaction log analysis data shows that geographic browse was the second most often used collection browse option in Opening History, with 18% of all collection browse instances.

Although all three participating historians mentioned that knowing the types of materials in a collection is important for them, none used the object browse option during the observation session. This finding slightly differs from the transaction log analysis findings, which show that object browse option is used in 17% of collection browse instances. However, while not knowingly initiating object browse, participants were also observed clicking on hyperlinked values in *Objects* collection metadata fields — an action which results in opening a respective collection browse results window.

The title browse option per se was not used in observation sessions. It was also the least used browse option based on the web log data, with only 5% of collection browse instances. However, all three participating historians browsed the lists of retrieved collections (arranged by collection title) after performing a search or faceted browse by subject, place, or hosting institution.

5.2.3 Value of Collection-Level Metadata

The interviews and observations of the three academic historians show that these scholars value collection-level metadata. All three participants reported that collection records are very helpful in providing information important for their research and teaching, although one of the participants pointed out that collection-level metadata is much less crucial than high-quality item-level metadata in aggregations of digital collections. This observation was supported by the fact that all three participants looked at the full versions of some of the collection-level metadata records that were retrieved as a result of their searches. Interestingly, the historian who reported that collection metadata was less important than item metadata, viewed full collection records, but never opened full item records in Opening History during the observation session. One respondent stated that collection metadata is more important than item-level metadata to him, and commented, "I don't necessarily need to see it [item metadata record], generally by that time [of looking at individual item] I know how it's been described as a collection. I know what I am looking for in general." While looking at 244 results retrieved by his item-level search, this historian commented on the value of collection metadata in organizing the item results: "if I am searching for something initially, this is too much information. I'd rather see it grouped by collection and have good metadata about the collection as a whole." For this researcher, the availability of item records appeared secondary to the contextual role of the collection information.

5.2.4 Important Collection Information and Metadata Fields

Participants named several collection metadata fields important in their information discovery in aggregations. Among the fields available in collection records in Opening History,

Subjects, *Size*, *Objects*, and *Geographic Coverage* were named by all three participants, while *Time Period* was named by only one participant. One of the two Native American history researchers also pointed out that it would be helpful to include tribal names in the *Subjects* fields, wherever applicable. *Provenance* was named by all three participants as a crucial kind of information about a digital collection. They noticed provenance information in free-text *Description* fields, but one of them pointed out that including a separate collection metadata field for provenance information would be helpful. The free-text *Description* field was named as important by all three participants, although one participant commented that in the case of a different, familiar digital collection, the *Description* field was "a little more flowery than it needs to be for my purposes." *Copyright* was named as an important collection metadata field to consult when doing research, but not for teaching-related information discovery. One researcher commented that he is not worried about copyright information "as long as it's there somewhere," i.e., on collection homepage.

5.2.5 Collection Metadata Display

When asked to compare the collection metadata displays in Opening History and American Memory, all three historians commented that the more structured approach taken by Opening History, which displays all collection metadata fields in a record, works better for them than the approach taken by American Memory. Here only the rich free-text Description field is displayed. The more structured records were preferred as more useful "because they do not require a lot of reading, while presenting information of interest (subject headings, object types, geographic coverage, etc.)." All three participants commented that they are more interested in structured and hyperlinked collection metadata, which supports browsing, for their research (e.g., "the thing that I like about this is that allows you to go use these links to search for similar types

of materials"), while just a long free-text *Description* field with photographs in it might be more useful for their teaching (e.g., "if I was looking for something to teach with, this would be very helpful, because I can read through it and get more sense of the background of this collection"). Figures 28 and 29 illustrate this comparison by showing two different collection metadata records and displays for the same collection — *Ansel Adams Photographs of Japanese American Internment at Manzanar* — in Opening History and American Memory. It was noted by one participant that overall, Opening History is better organized for subject discovery, while American Memory is more useful for known-item discovery ("if you know a particular collection"), but also that "you can find a lot more things" in Opening History.

5.2.6 Other Considerations

Two participants of this study shared their observations on current digitization practices. One of them commented that the types of historical materials most often selected for digitization do not match the object types that are useful in study in his research area (Japanese American history): "they don't tend to digitize the things which I need the most, which is primary documents, government documents or manuscript collections; things like this are usually the last things that archives tend to digitize." He explained that archives primarily digitize photographs and other images, and that "for historians that deal with images this is helpful but for historians that don't deal with images — unfortunately not." Another scholar complained that digital collections of historical content are created haphazardly, without giving much thought to which materials might be useful, often without consulting history faculty and students. He greatly prefers focused or specialized collections to the collections with a broad spectrum of coverage. One historian made an observation about the redundancy in digitization efforts: he noticed that

particular historical materials digitized by one state's historical society were also digitized and included in a digital collection in another state.

One of the study participants voiced questions about the values in collection metadata fields in the Opening History aggregation. In particular, she asked how the decision is made about what audience terms to use to populate the *Audience* field. The scholar commented that she would be looking for collections with "Researchers" as a value but would be reluctant to further explore a collection with "K-12 students" as one of the values. She was also confused by the *Interaction with Collection*, unsure what was meant by the name of the field, as well as its values (e.g., "search", "browse"), in the context of collection description. Another historian asked how the subject headings used in the subject browse option (*GEM Subjects*) are selected. He commented that "older historians might struggle a little bit just understanding what all of this [some metadata field names and values] means," although collection metadata records are "kind of like a library catalog." A particular concern about GEM subject headings was voiced: "I can tell that it was originally created for educators, because 'social studies' is not a category that historians would recognize."

One participant commented on the uneven quality of item metadata records and maintained that minimal item-level metadata (e.g., without creator or provenance information) makes interesting content not useful for scholars. He praised some digital collections "with thorough and detailed item-level descriptions" and pointed out that adhering to the standards in item-level metadata is needed for digital collections to be more useful to historians.

Participants expressed interest in hyperlinking the values in collection metadata fields just as in the library catalogs, to allow for more focused subject browsing. Two respondents

commented on the usefulness of the hotlinks in the *Subjects* collection metadata fields,; one asked why all of the subject headings were not hyperlinked. Two participants wondered why the values in the *Time Period* field were not clickable, unlike the values in similar fields (e.g., *Geographic Coverage*, *Subjects*, and *Objects*) to help in browsing by time period. However, they also were confused as to where such links should lead: to the list of all digital collections that cover a certain time period, or to the list of items in a given collection that deal only with this time period.

One respondent asked about the principle used for ranking of search results in Opening History and whether it was roughly chronological or best match. She commented that the latter was her preferred method and that it would be nice if Opening History "like ProQuest, allow[ed] different ways of sorting" the search results.

5.3 User Interactions Findings in Relation to Prior Work on Scholarly Searching

Similar to Zhang and Salaba's (2007b) research, this study of scholarly users revealed that system functions supporting user tasks involved in resource discovery by subject — subject browse, keyword search, free-text *Description* fields with content abstracts — are helpful in searching and resource discovery. Both transaction log analysis results and interview and observation results support previous studies showing that browsing remains a prevalent form of user interaction (Borgman 1996; Hildreth, 1995; Ellis & Oldman, 2005). In particular, subject and geographic browse were found to be widely used. Although the interview and observation did not reveal a preference for any specific way of interacting with aggregations, the findings of

transaction log analysis disagree with Buchanan et al. (2005) observation that humanists and social scientists prefer searching to browsing in digital library environment.

Results of earlier studies, which found that users of online catalogs tend to search more often by keyword than use any other type of search (e.g., Fidel, 1988, 1992; Curl, 1995; Hildreth, 1997; Muddamalle, 1998; Spink & Jansen, 2004) are supported by the transaction log analysis, which shows that in most cases (69% of item-level searches) the users of aggregations ignore advanced search options and use the simple keyword search option. However, results of this study reveal much higher level of advanced searching in aggregations (30.8%of item-level searches and 8.97% of collection-level searches) compared to the findings of other studies. This finding may be attributed to the possible differences between the presumably mainly scholarly audience of aggregations of cultural heritage digital collections, and the wider audience of Web users. Another possible explanation may be derived from the Jansen and Spink (2006) observation that because different searching contexts influence information searching behavior, it is often impossible to apply results from studies of one particular Web search engine to another Web search engine.

This research confirms Bates (1996) taxonomy of key search query term types that appear in the searches of art historians — names of individuals, geographical names, chronological terms, and discipline terms — by showing that query term categories in an aggregation of digital collections aimed at scholarly historians often include FRBR individual *persons*, *places*, *events*, and *concepts*. However, this study also found additional search categories performed in aggregation: *object*, *corporate body*, *work*/*collection*, *ethnic group*, and *class of persons* — findings which were not indicated in Bates' earlier study.

Both Bates (1996) and this study's observations about the prominence of geographic search differ from Kipp's (2006) findings that users pay less attention than indexers to geographic location in assigning the subject tags to journal articles on *CiteULike*. This contrast may highlight different approaches in active tagging than in the more passive use of available tags or headings assigned by the indexers. On the other hand, this difference might indicate information seeking behavior patterns that are unique/specific to the particular user communities (art historians in Bates study and scholarly U.S. historians in this research), or a particular information context (aggregations of digital collections) and may not be generalizable to a wider population.

The findings of the observation sessions support suggestions made by developers and researchers of aggregations (e.g., Foulonneau et al., 2005; Harum, 2008) that providing links between item-level records and relevant collection-level records in aggregations can facilitate browsing behavior familiar to humanities scholars in general, and historians in particular. Historians who participated in the study were found to examine collection metadata records in the course of their interaction with Opening History and to follow the links to item metadata from collection metadata records, as well as to go from the item search results to collection descriptions of digital collections containing the items of interest.

Other findings of this dissertation research are similar to results of the earlier research into the effect of domain knowledge on information seeking behavior. In the early 1990s, online catalog users' domain knowledge levels were found to influence information seeking behavior and outcomes of the search in online catalogs (Allen, 1991), while domain experts were found to have clear expectations for both the answer and the context in which it would appear (Marchionini et al., 1993). The academic historians observed and interviewed in this research

were all domain experts both in their general field (history) and their specific subject areas (Native American history, Japanese American history). They also have clear collection expectations for familiar aggregations and digital collections where the digital content relevant to their research could be found. Pennanen, Serola, and Vakkari (2003) observed that academic users with higher domain knowledge on their research topic use wider and more specific vocabulary in their subject search, while Zhang, Anghelescu, and Yuan's study (2005) suggested that as the level of domain knowledge increases, users tend to do more searches and to use more terms in queries. While domain expertise information cannot be associated with the log data, academic researchers are the primary target audience for Opening History. Interestingly, the average search query length in the Opening History web logs is rather low(1.5 words per query), but the queries tend to be quite specific, whether they consist of several words (e.g., "boston city directory 1911", "Olasee Davis", "ship plans" etc.) or just one specific term (e.g., "amphibians", "edgewater", "tipi" etc.). In observation sessions as part of this study, the historians relied on rather specific search terms (e.g., "pueblo", "Crow Indians", "war relocation authority") and used multi-word queries more often that single-word queries.

Results of early catalog use studies (summarized by Krikelas, 1972), which looked into the use of different description areas in traditional card catalogs, offer an interesting point of comparison to the findings of this research. In card catalogs, patrons reported heavy use of subject headings, while fields like content note tended to be consulted less often, and size information was rarely used by library patrons. More than forty years later, interview shows that various kinds of subject headings (topical, geographical, and temporal) are still of great value to the users of aggregations of digital collections,, who are also very interested in two areas — size information and free-text *Description* fields (roughly equivalent to content note in a catalog) —

which were less often or rarely used by library patrons in traditional card catalogs. The log data collected and analyzed in this study supports the interview and observation findings by showing that subject browse, which is based on subject headings, is a heavily used form of interaction with an aggregation of digital collection. The search in aggregation also demonstrates that a high proportion of user search terms is satisfied through the subject headings in *Subjects* collection-level metadata fields.

5.4 Summary

Overall, item-level user interactions with Opening History — both search and browse — were found to occur more often than collection-level user interactions. Browse — both collection-level and item-level — was found to be initiated more often than search. Among the collection-level browse queries, subject browse was used most often, followed by geographic and object type browse.

Most collection-level search queries in Opening History were found to fall within FRBR Group 3 categories: *concept*, *object*, and *place*, and thus can be considered subject searches. Collection searches observed in Opening History differ somewhat from item-level searches. More *object*, *concept*, and *corporate body* searches and fewer *person* searches were observed at the collection level than at the item level.

When the actual user collection searches from the web log data were repeated in the Opening History, they were found to be most often satisfied by *Description* and/or the *Subjects* collection metadata fields. Moreover, a significant proportion of collections would not be retrieved in collection-level searches in without controlled-vocabulary subject metadata fields

such as *Temporal Coverage*, *Geographic Coverage*, *Subjects*, and *Objects* (42% overall), and free-text metadata (*Description* field — 21%).

Interviews and observations of academic historians revealed that scholarly users value collection metadata. This finding is supported by the results of transaction log analysis, which shows that collection metadata records were viewed more often than any kind of collection search or browse was initiated. Historians expected to see information about provenance, collection size, types of objects, subjects, geographic coverage, and temporal coverage in collection-level metadata. The structured display of collection metadata in Opening History was found to be more useful for historians (especially for their research needs) than the alternative approach taken by American Memory, which displays only the free-text *Description* metadata field and uses the rest of its rich collection metadata behind-the-scenes to support information retrieval. Interviews and observations also collected feedback from scholarly historians on related topics which were outside the scope of this study, including metadata quality and digitization practices.

5.5 Figures and Tables

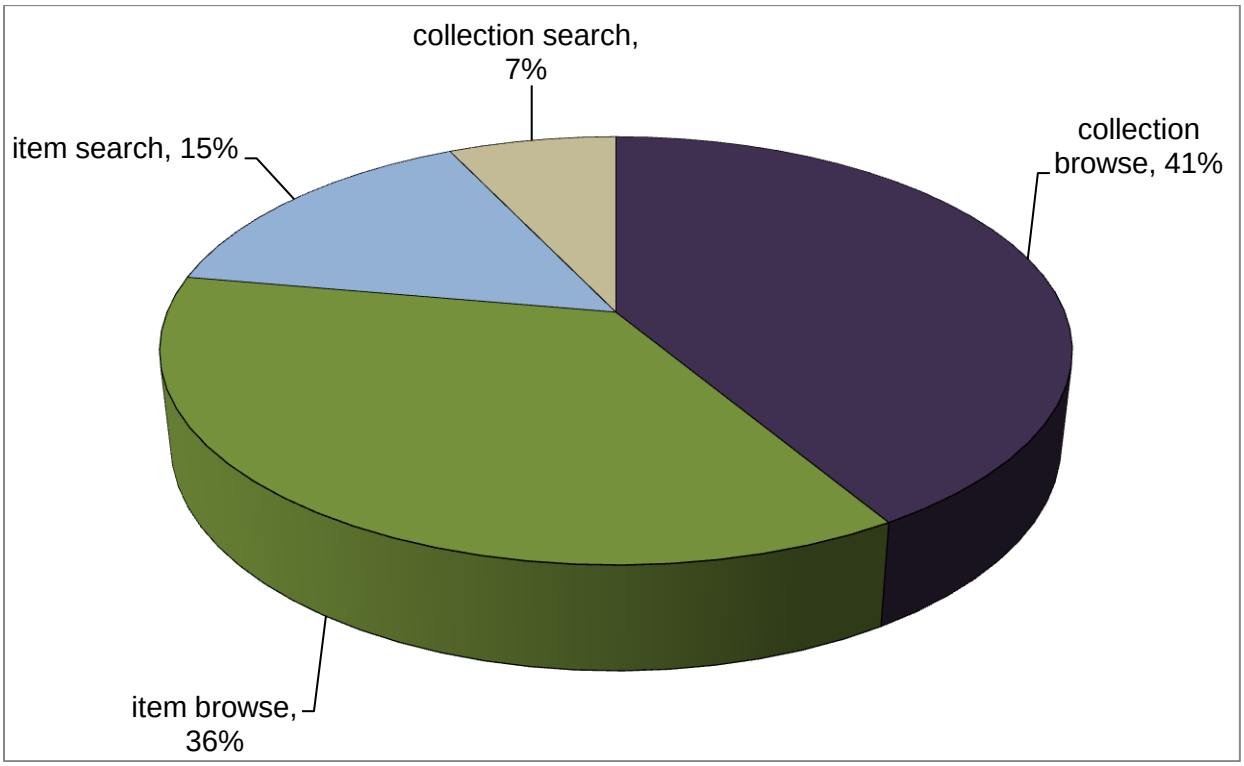


Figure 19. Search and browse in Opening History: pie chart

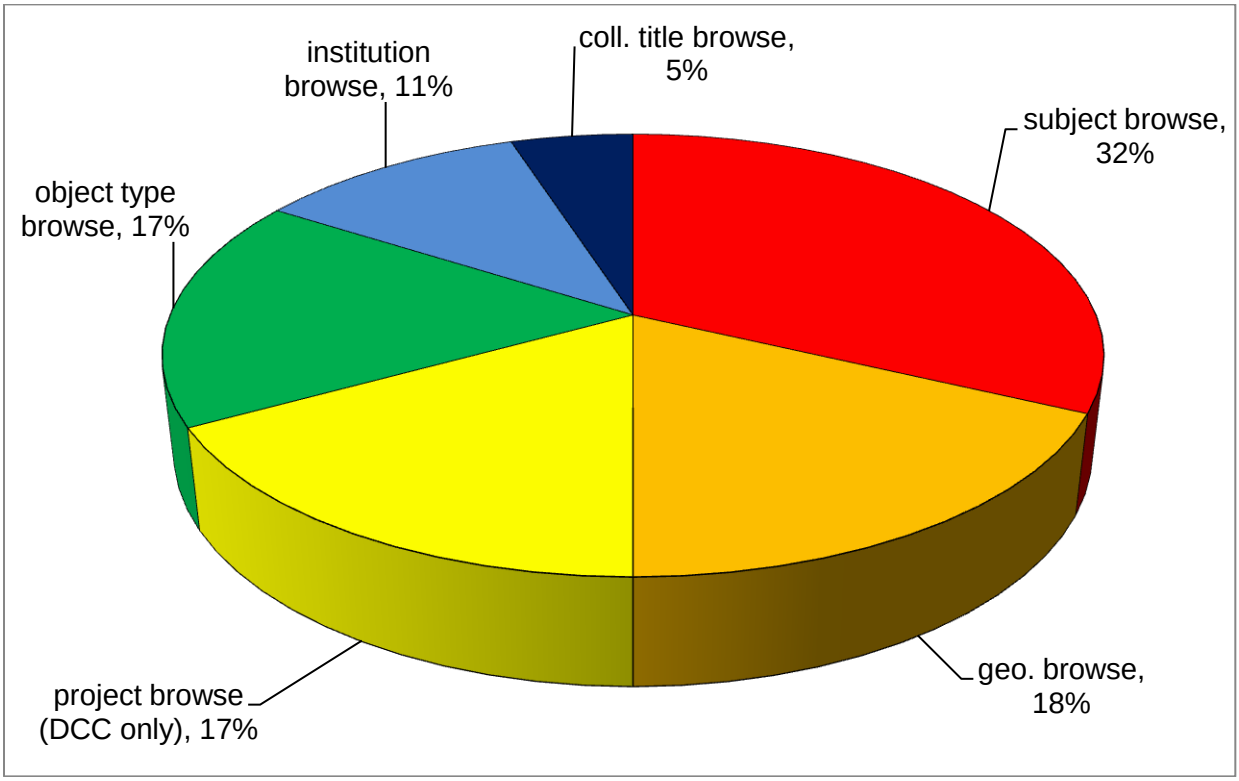


Figure 20. Collection browse types in Opening History: pie chart

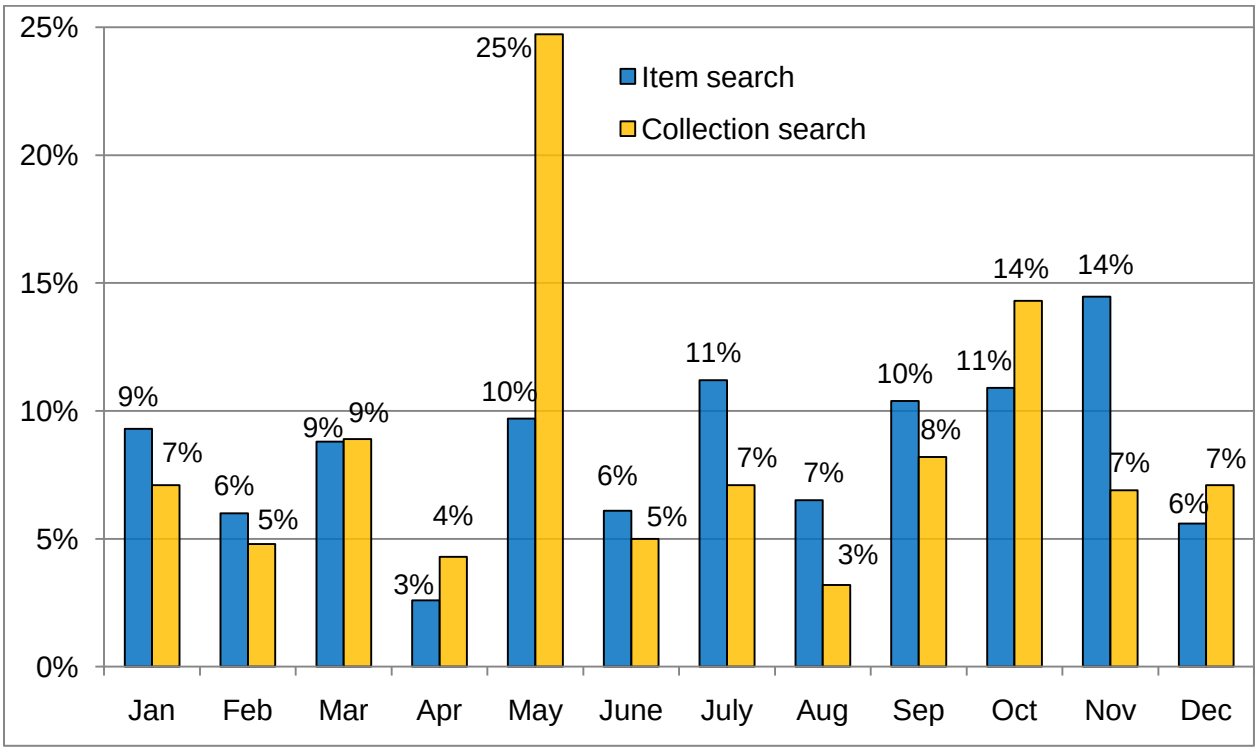


Figure 21. Distribution of collection and item searches in Opening History by month

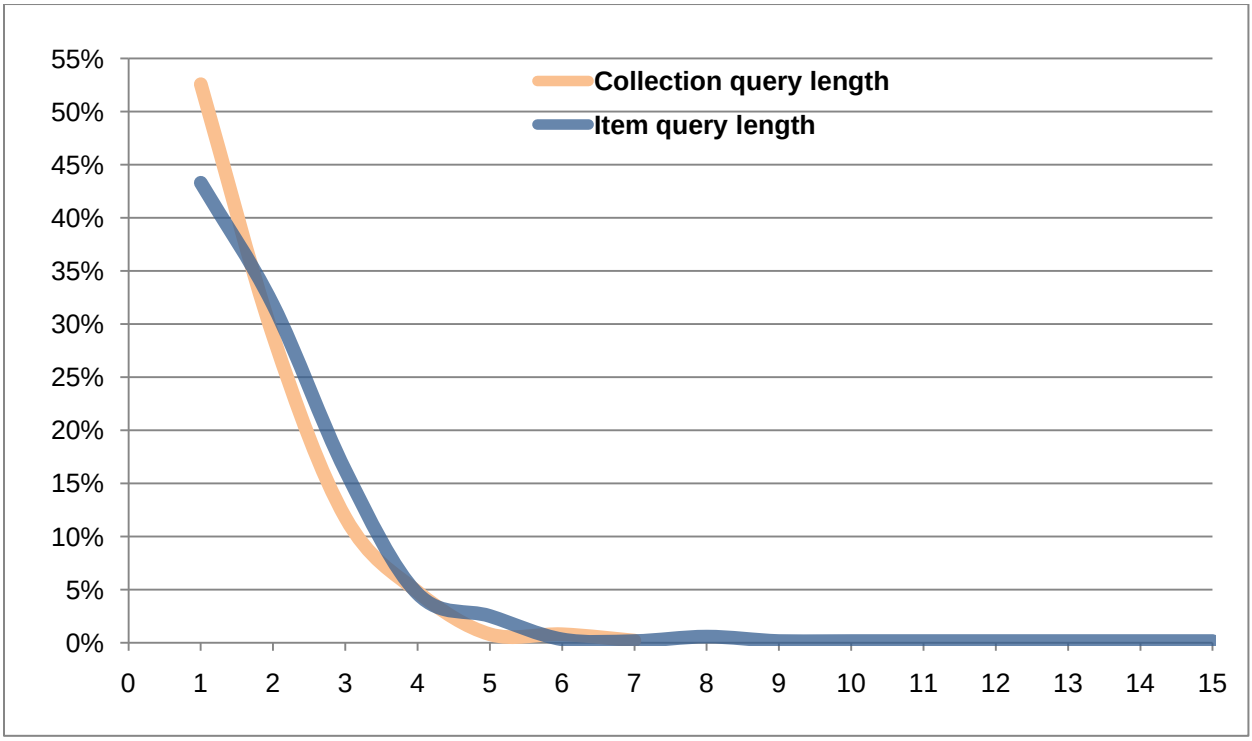


Figure 22. Distribution of collection and item search query lengths

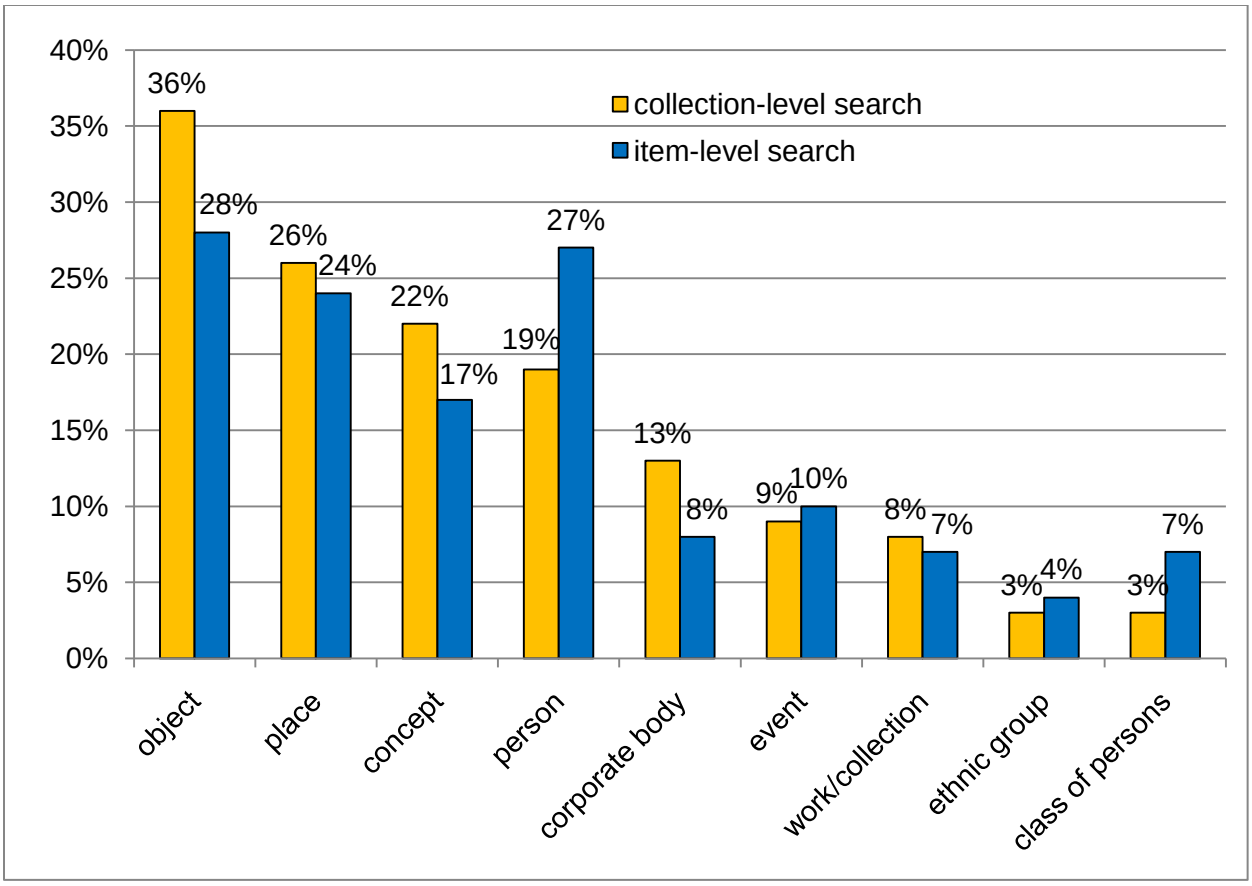


Figure 23. Distribution of user searches in Opening History by FRBR-based search categories

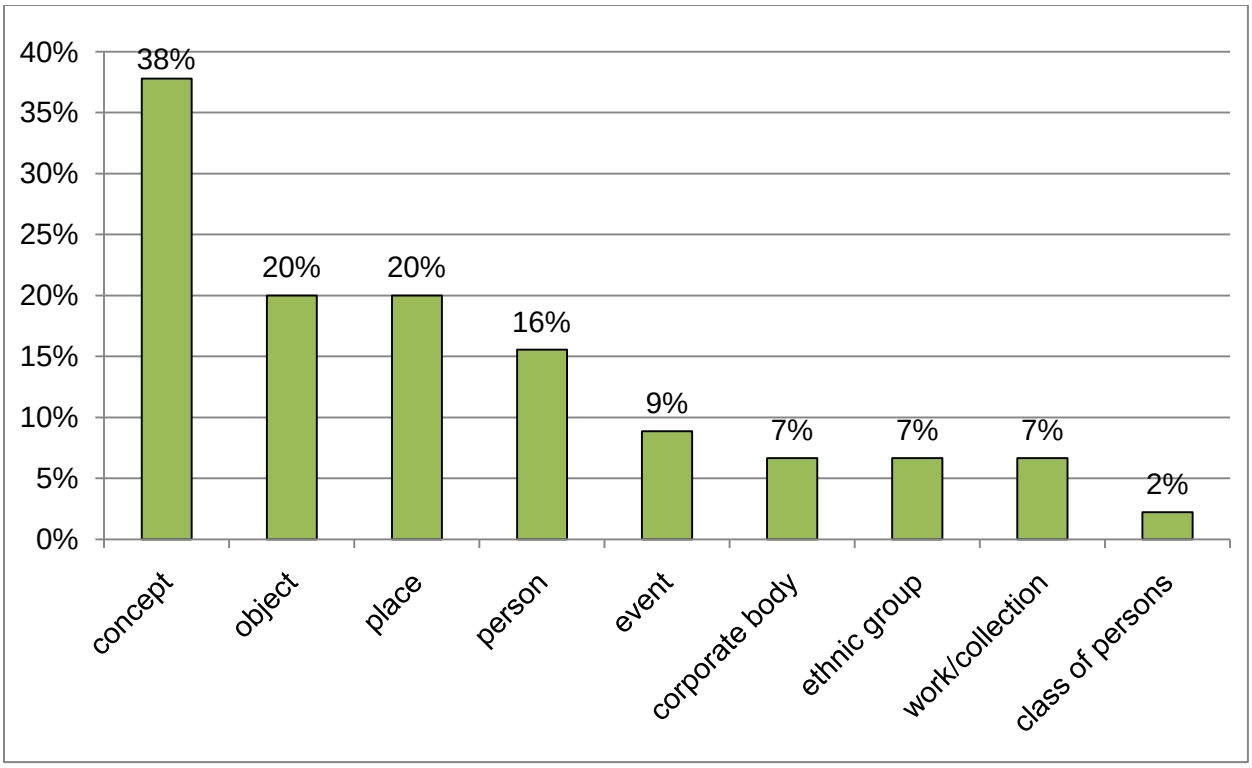


Figure 24. Distribution of user searches in Opening History by FRBR-based search categories: overlap between collection and item searches

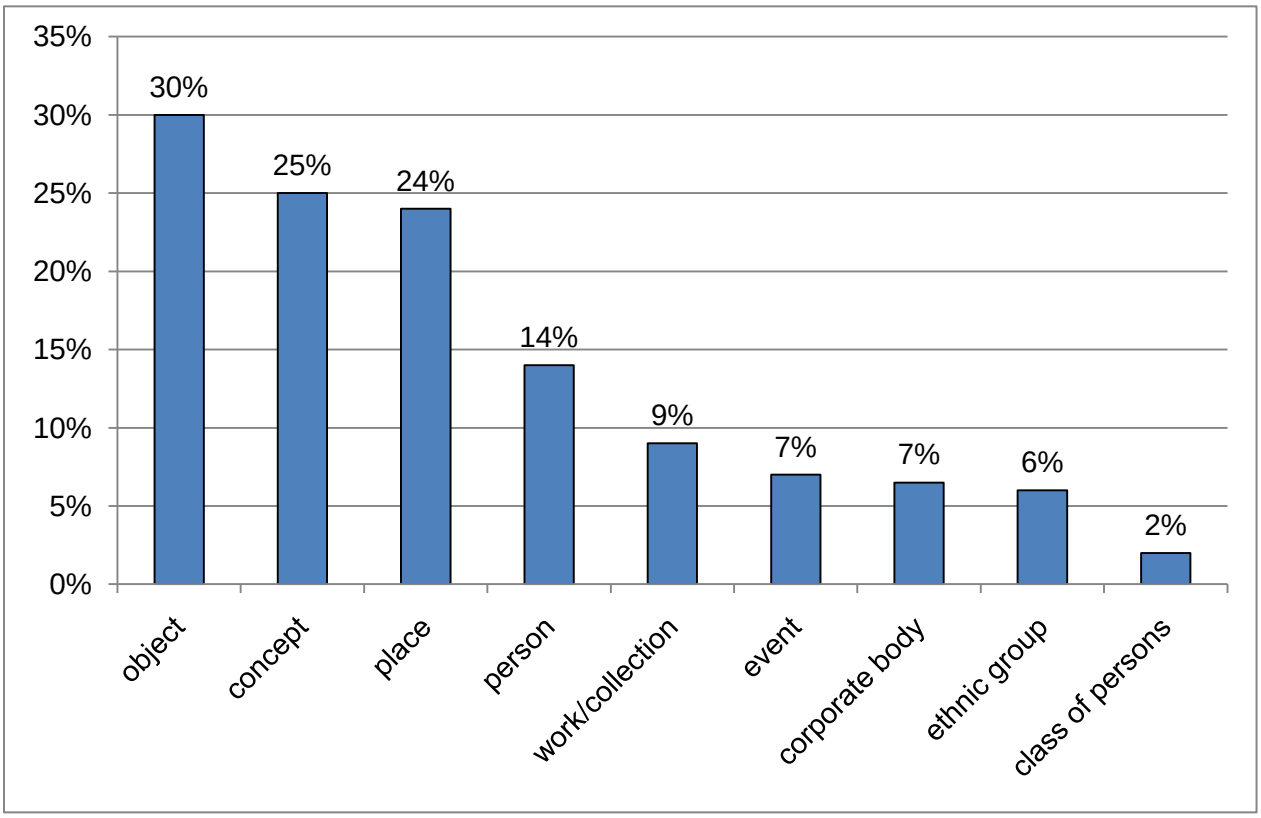


Figure 25. Distribution of successful collection-level user searches in Opening History by FRBR-based search categories

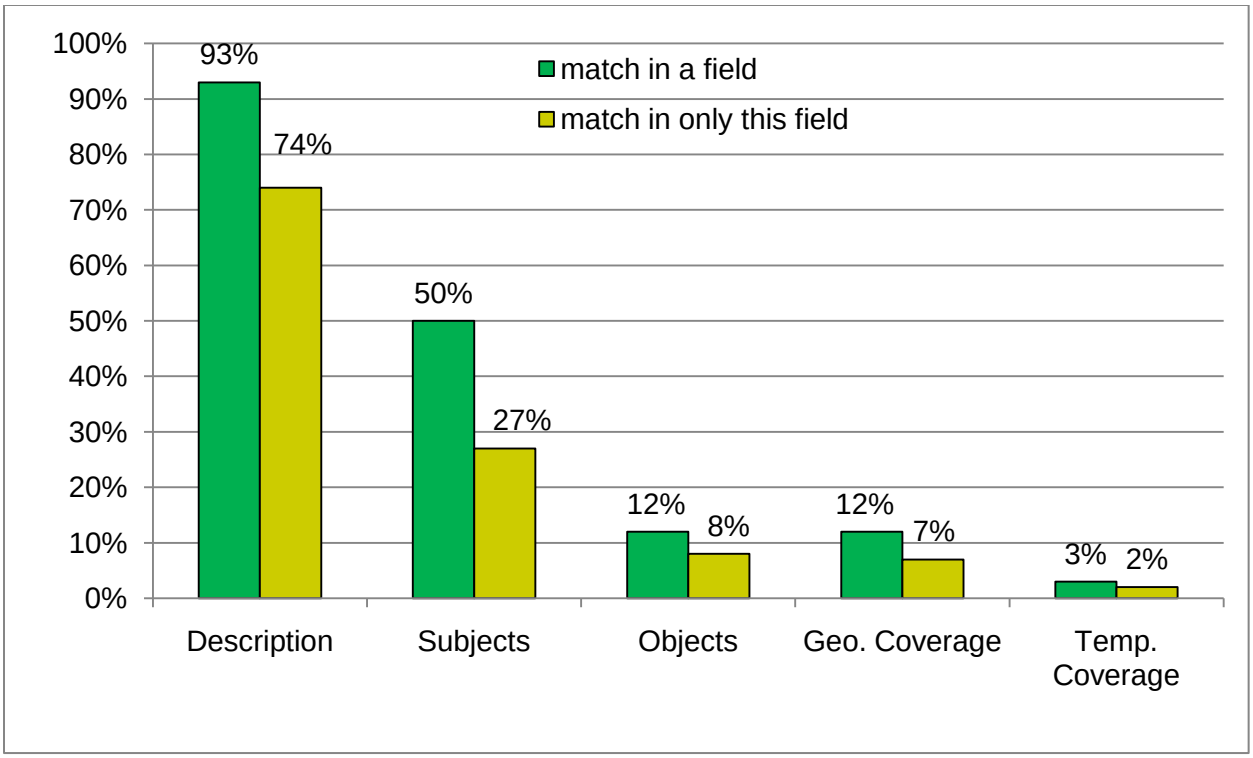


Figure 26. Subject-specific collection metadata fields matching user search queries

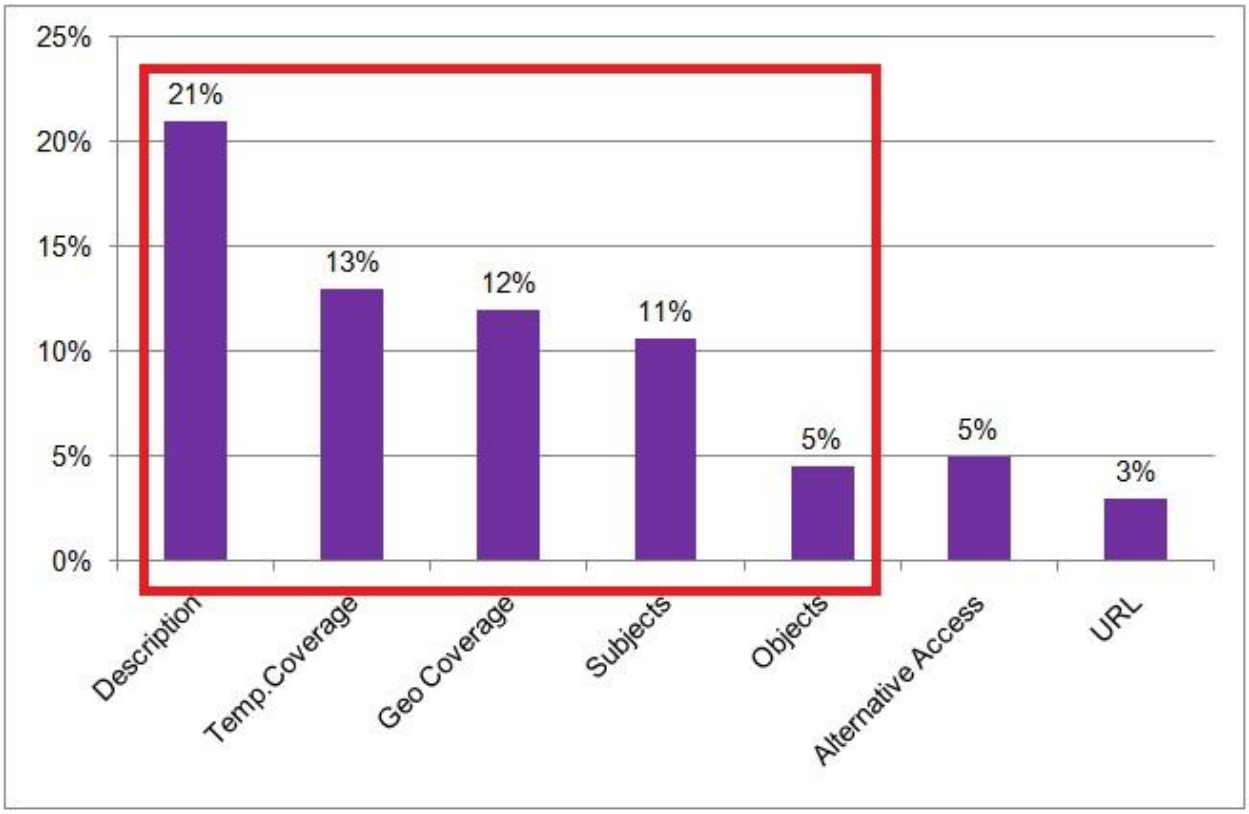
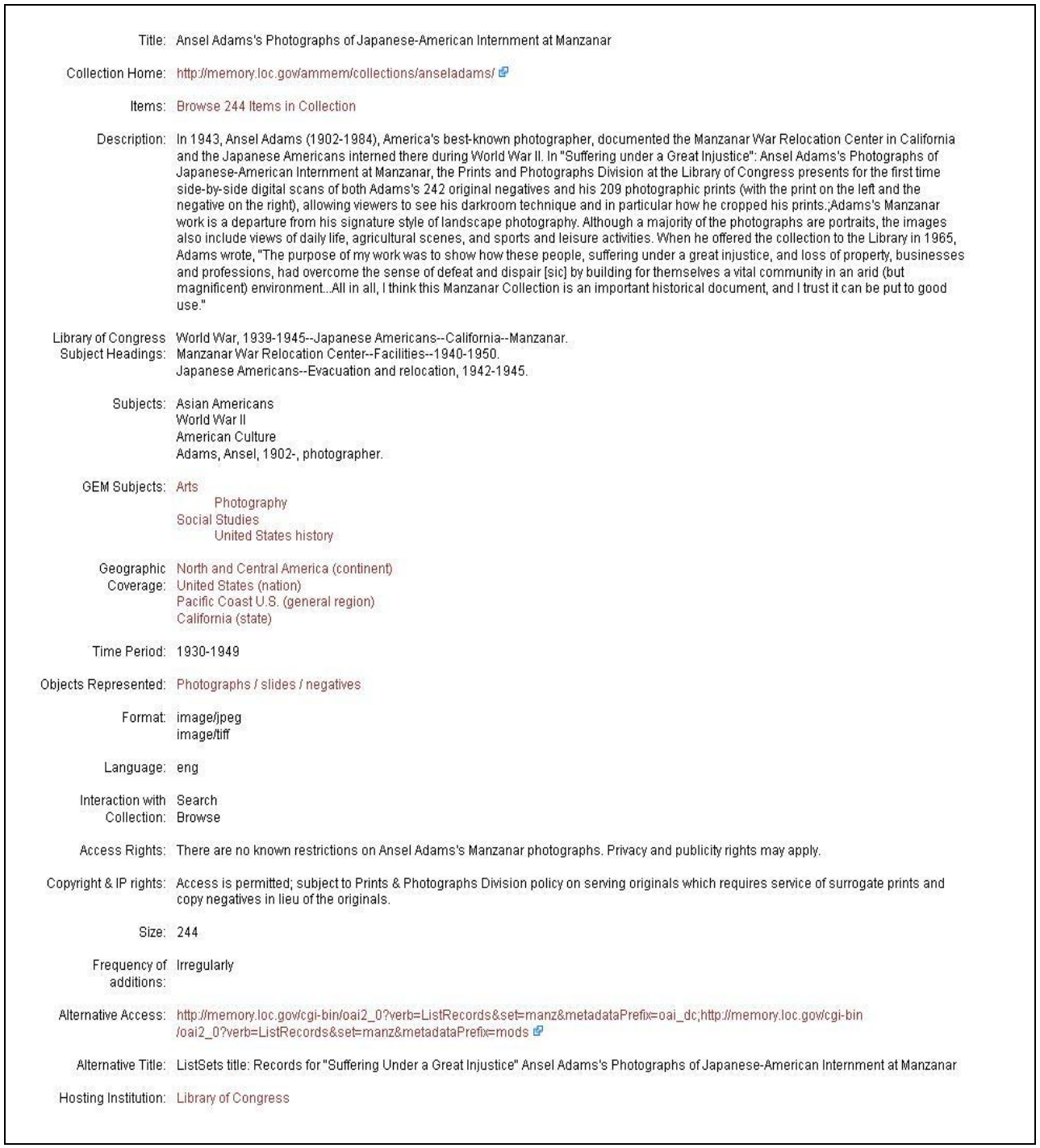


Figure 27. Percentage of collection records that would not be retrieved in collection search if certain collection metadata fields were absent



Title:	Ansel Adams's Photographs of Japanese-American Internment at Manzanar
Collection Home:	http://memory.loc.gov/ammem/collections/anseladams/
Items:	Browse 244 Items in Collection
Description:	In 1943, Ansel Adams (1902-1984), America's best-known photographer, documented the Manzanar War Relocation Center in California and the Japanese Americans interned there during World War II. In "Suffering under a Great Injustice": Ansel Adams's Photographs of Japanese-American Internment at Manzanar, the Prints and Photographs Division at the Library of Congress presents for the first time side-by-side digital scans of both Adams's 242 original negatives and his 209 photographic prints (with the print on the left and the negative on the right), allowing viewers to see his darkroom technique and in particular how he cropped his prints.;Adams's Manzanar work is a departure from his signature style of landscape photography. Although a majority of the photographs are portraits, the images also include views of daily life, agricultural scenes, and sports and leisure activities. When he offered the collection to the Library in 1965, Adams wrote, "The purpose of my work was to show how these people, suffering under a great injustice, and loss of property, businesses and professions, had overcome the sense of defeat and dispair [sic] by building for themselves a vital community in an arid (but magnificent) environment...All in all, I think this Manzanar Collection is an important historical document, and I trust it can be put to good use."
Library of Congress Subject Headings:	World War, 1939-1945--Japanese Americans--California--Manzanar. Manzanar War Relocation Center--Facilities--1940-1950. Japanese Americans--Evacuation and relocation, 1942-1945.
Subjects:	Asian Americans World War II American Culture Adams, Ansel, 1902-, photographer.
GEM Subjects:	Arts Photography Social Studies United States history
Geographic Coverage:	North and Central America (continent) United States (nation) Pacific Coast U.S. (general region) California (state)
Time Period:	1930-1949
Objects Represented:	Photographs / slides / negatives
Format:	image/jpeg image/tiff
Language:	eng
Interaction with Collection:	Search Browse
Access Rights:	There are no known restrictions on Ansel Adams's Manzanar photographs. Privacy and publicity rights may apply.
Copyright & IP rights:	Access is permitted; subject to Prints & Photographs Division policy on serving originals which requires service of surrogate prints and copy negatives in lieu of the originals.
Size:	244
Frequency of additions:	Irregularly
Alternative Access:	http://memory.loc.gov/cgi-bin/oai2_0?verb=ListRecords&set=manz&metadataPrefix=oai_dc;http://memory.loc.gov/cgi-bin/oai2_0?verb=ListRecords&set=manz&metadataPrefix=mods
Alternative Title:	ListSets title: Records for "Suffering Under a Great Injustice" Ansel Adams's Photographs of Japanese-American Internment at Manzanar
Hosting Institution:	Library of Congress

Figure 28. Collection metadata display example: Opening History

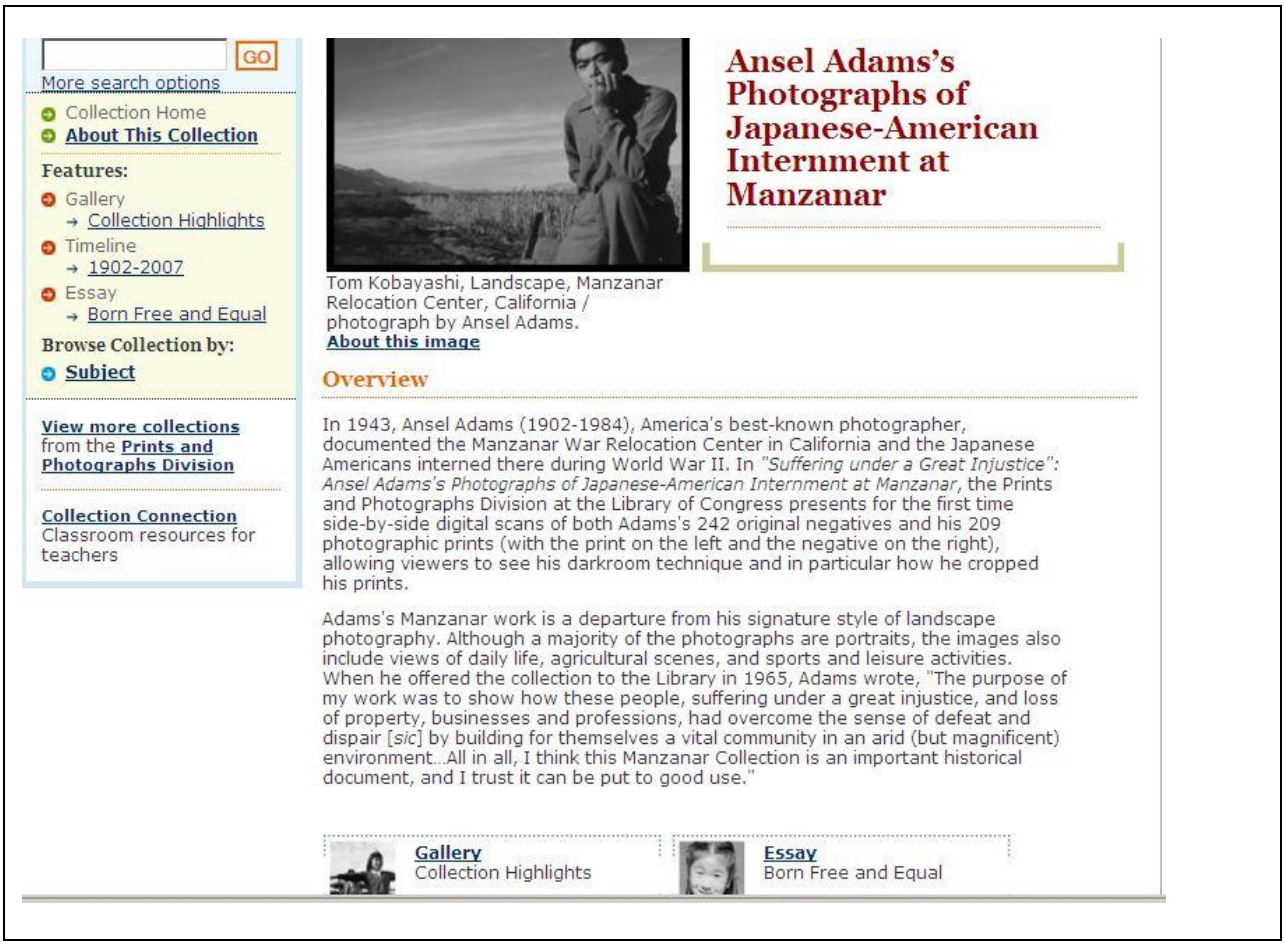


Figure 29. Collection metadata display example: American Memory

Types of user interaction	Number of pageviews in the sample
collection browse	4,939
item browse	4,388
item search	1,860
collection search	880

Table 10. Search and browse in Opening History: pageview statistics

Collection browse type	Mean	Median	Variance	Standard deviation
Subject browse	11	1	17.07	4.13
Geographic browse	2.29	1	7.59	2.76
Project browse	2.03	2	2.25	1.50
Object browse	2.64	1	9.57	3.09
Institution browse	1.77	1	2.59	1.61
Title browse	7	7	50	7.07

Table 11. Collection browse types in Opening History: variability measures

User interactions	Pageviews
Viewing collection metadata records	1760
Viewing item metadata records	368
Collection browse:	2939
subject browse	953
geographic browse	533
project browse	502
object type browse	487
institution browse	311
collection title browse	153
Item browse	4388
Collection search	880
Item search	1860

Table 12. Collection and item metadata records pageviews in comparison with other user interactions

	Mean	Median	Variance	Standard deviation
Collection search				
Length	1.75	1	1.03	1.01
Frequency	1.54	1	1.31	1.14
Item search query:				
Length	1.99	2	1.69	1.30
Frequency	1.89	1	14.94	3.87

Table 13. Search query length and frequency: variability measures

Collection metadata field	% of searches retrieving 1 or more collection records with a match in a field	% of searches retrieving 1 or more collection records with a match ONLY in this field
Description	93.0%	74.0%
Subjects	50.0%	27.0%
Title	48.4%	4.9%
URL	18.5%	2.2%
Copyright & IP rights	15.2%	2.7%
Objects	12.0%	8.0%
Geographic Coverage	12.0%	7.0%
Alternative title	11.4%	1.6%
Notes	8.7%	3.8%
Alternative access	6.5%	1.1%
Size	6.0%	1.1%
Contributing institution	6.0%	0.0%
Access rights	4.9%	0.5%
Creator of collection	4.0%	0.0%
Temporal Coverage	3.0%	2.0%
Audience	2.2%	1.6%
Interaction	1.6%	1.1%
Supplementary materials	1.6%	0.5%
Hosting institution	1.6%	0.0%
Format	0.5%	0.5%
Provenance	0.5%	0.5%
Collection development policy	0.5%	0.5%
Metadata schema used	0.5%	0.5%

Table 14. Collection metadata fields with matches to user search terms

Other areas of collection descriptions with the matches to user search terms	% of searches retrieving 1 or more collection records with a match in an area
Complementary digital collection	6.0%
Sub-collection	2.1%
Associated physical collection	1.1%
Parent collection	1.1%
Associated project	1.1%

Table 15. Other areas of collection description with matches to user search terms

Chapter 6. Conclusions, Implications, Limitations, and Future Research

This chapter summarizes the most important results of the study, draws conclusions, discusses the implications and limitations of the study, and provides an overview of future research.

6.1 Conclusions and Implications

This dissertation research sought answers to the question: ***How does collection-level metadata mediate scholarly subject access to aggregated digital collections?*** More specifically, this study examined how subjects are represented in collection-level metadata in aggregations of digital collections, as well as how scholarly historians interact with aggregations of digital collections, and what role collection-level metadata (free-text and structured) and its richness play in such interaction. Using three research methods: comparative content analysis of collection-level metadata in three large-scale aggregations of cultural heritage digital collections (Opening History, American Memory, and The European Library), transaction log analysis of user interactions with Opening History aggregation, and interview and observations of academic historians interacting with two aggregations of cultural heritage digital collections (Opening History and American Memory), this study arrived at the conclusions summarized below in sections 6.1.1, 6.1.2, 6.1.3, and 6.1.4. Table 16 relates the research questions asked in this study with the answers found. Sections 6.1.5 and 6.1.6 summarize theoretical contributions of this research and present practical implications for development of collection metadata in aggregations of digital collections.

6.1.1 Overall Value of Collection Metadata and Its Display to Scholarly Users of Cultural Heritage Aggregations

The value of collection-level metadata and representation was evident to the academic historians studied in interviews and observations Participants of the observation viewed at least one collection metadata record during their interaction with an aggregation; most often they viewed collection metadata records after conducting a collection search or browse, but sometimes also after item search. Respondents stated their preference for how information is organized at the collection level and found collection record information sufficient for gaining an understanding of the content of interest. For at least one participant the availability of item records appeared secondary to the contextual role of the collection information. This finding is supported by the transaction log analysis results, which shows a high level of engagement with collection metadata records, with the total page views for collections more than 4 times greater than item page views. Transaction log analysis also shows that collections metadata was viewed approximately as often as the two most widely found in the logs types of collection queries — collection search and collection-level subject browse — taken together were initiated. This allows to conclude that viewing collection metadata is an integral component of any collection-level interaction, as well as some item-level interactions.

The structured display of collection metadata in Opening History was found to be more useful for historians participating in the interview and observation sessions (especially for their research needs) than the alternative approach taken by many other aggregations, including American Memory, which displays to the end-user only the free-text *Description* metadata field and uses the rest of its rich collection metadata behind-the-scenes to support information

retrieval. Interviews with historians also suggested that the structured collection metadata display works more effectively for subject discovery, while the free-text *Description* alone is suitable for known-item information search (at least in the familiar American Memory aggregation which is not currently growing). Developers of the large and steadily growing aggregations should take the influence of collection metadata structure and display on subject access into consideration when making their decisions.

6.1.2 Richness of Collection Metadata

Subject-based resource discovery is significantly influenced by collection-level metadata richness. The collection-level metadata richness includes such components as 1) describing collection's subject matter with mutually complementary values in different metadata fields and 2) variety of collection properties/characteristics encoded in the free-text *Description* field.

A total of 19 different collection characteristics found in free-text *Description* fields across the three aggregations: Opening History, American Memory, and European Library. The free-text *Description* collection metadata field was found to contain on average 6 different collection properties or characteristics. Types and genres of objects in a digital collection, topical subjects, geographic and temporal coverage were found to be the most consistently represented collection characteristics. Five collection characteristics were found only in the free-text *Description* fields and no other collection metadata field: the creator of items in a digital collection, the provenance, the uniqueness, importance, and comprehensiveness of content in a digital collection.

The information found in different collection metadata fields is often mutually complementing. The assumption, based on the pilot studies, that *Description* field would often

complement other metadata fields, was supported by this study's findings. It was also observed in this study that information in other collection metadata fields complements information in *Description* field almost as often, sometimes even more often (as in case with *Geographic Coverage* field). The cases of two-way complementarity between the values encoded in two metadata fields were observed less often than one-way complementarity, mostly between *Description* and *Subjects* fields and between *Description* and *Geographic Coverage* fields. Little redundancy between the values in different collection metadata fields was observed.

These findings demonstrate the richness of collection metadata records in large-scale aggregations of digital collections. Results of this study indicate that encoding of mutually complementary subject-specific information in free-text and controlled-vocabulary metadata fields is already being recognized as a benchmark in crafting rich collection-level metadata in aggregations. In addition to subject-specific information, the emerging best practices in collection-level description observed in this study suggest enriching *Description* fields by encoding a variety of other collection characteristics such as title, size, collection development policy, copyright information, provenance, intended audience, navigation and functionality, language of items in collection, frequency of additions, participating or contributing institutions, funding sources, and especially the characteristics for which no specialized collection metadata fields exist: collection strengths (importance, uniqueness, and comprehensiveness) and creators of items in collection.

6.1.3 Role of Collection Metadata in User Interactions with Aggregations of Digital Collections

Overall, collection-level user interactions — search and browse — were found to occur less often than item-level user interactions recorded by the Opening History web logs. Browse — both collection-level and item-level — was found to be initiated in Opening History more often than search.

Among the collection-level browse queries, subject browse was used the most often, followed by geographic and object type browse. This observation agrees with the studies of historians' information seeking behavior in aggregations and other full-text digital library environments (Harum, 2008; Wu & Chen, 2007), which found that historians prioritize geographic browse capability and value subject browse, and with the findings of interviews that show historians' interest in object type information. Because the browse capability usually relies on controlled-vocabulary values encoded in respective specialized collection metadata fields, it is desirable that cultural heritage aggregations systematically apply specialized *Subjects*, *Geographic Coverage*, and *Object* collection metadata fields. Importance of these controlled-vocabulary subject metadata fields is underlined by the finding that a significant proportion of digital collections (42% overall) would not be retrieved in collection searches in Opening History without *Subjects*, *Geographic Coverage*, *Temporal Coverage*, and *Objects* fields. Therefore these collection metadata fields are crucial for collection-level subject access and should be applied more consistently in collection metadata in aggregations.[56]

[56] At the moment, only one aggregation among the three studied in this research (Opening History) consistently applies all five of these collection metadata fields in 100% of its collection records, while the two others (American Memory and European Library) are less consistent in applying *Temporal Coverage* (62%-85%) and *Geographic Coverage* (81%-92%) fields, and one (European Library) is very inconsistent in applying *Object*s field (41%).

In the interviews, historians reported that they expect to see types of objects, subjects, geographic and temporal coverage information about digital collection in collection metadata. When the actual user searches from the web log data were repeated in Opening History aggregation, the free-text *Description* collection metadata fields were found to provide the vast majority of the matches to user search terms. Moreover, a significant proportion of collections would not be retrieved in collection searches in Opening History without a free-text *Description* field (21%), more than any other metadata field taken alone, which suggests that his field is the most important for collection-level subject access.

Most collection-level search queries in Opening History were found to fall within FRBR Group 3 search categories: *object*, *place*, and *concept*. This finding provides insights into the kinds of information that should be present in collection-level metadata records to facilitate subject access to digital collections. While the most widely observed collection search categories (*object*, *concept*, and *place*) closely correlate with the three of the four most consistently represented collection characteristics in free-text *Description* field (types and genres of objects in a digital collection, topical subjects, and geographic coverage), other regularly occurring searches such as *corporate body* are not currently sufficiently represented in the free-text *Description* fields. The provenance and hosting/contributing institution information, which usually contains corporate body names, was found in under a third of collection-level *Description* fields in this study, moreover names of the provenance-related corporate bodies are not encoded anywhere else in the records.

Interview participants emphasized two kinds of collection metadata as important for their information needs: collection provenance and collection size. Since the *Provenance* field is not utilized by any of the collection metadata records analyzed in this study, free-text *Description*

field remains the only place in the record that can provide matches to the users' provenance-related search terms. It is therefore recommended that to better serve the academic historians' needs, provenance information in cultural heritage aggregations should be encoded in *Description* fields much more consistently than it is now (in 33%-37% of collection metadata records in three aggregations). Similarly, in two out of three aggregations analyzed in this study — Opening History and The European Library — size information was encoded inconsistently (in 28% and 37% of the *Description* fields respectively). While the separate *Size* collection metadata field is included and displayed to the user in most of the records in Opening History, The European Library does not have the specialized *Size* field in its collection metadata schema, which makes including this information in the *Description* field more important for this aggregation.

6.1.4 FRBR Model as a Conceptual and Analytical Framework for Studying Collection-Level Subject Access

This study used the Functional Requirements for Bibliographic References (FRBR) set of entities to analyze collection-level search queries performed in Opening History aggregation. The FRBR model proved useful as a framework for understanding collection-level subject access. It was found that both FRBR Group 3 (*concept*, *object*, *event*, and *place*) and Group 2 entities (*person*, and *corporate body*), as well as one of the Group 1 entities — individual *work* — are widely represented among the collection-level searches in aggregation of digital collections. The searches for specific named digital *collections* were also observed. Although *collection* entity is not explicitly represented in FRBR model, collections might be accommodated with the "is part of" relation between *works* of different levels.

Additional search categories emerged that are important for understanding the subject access in aggregations of digital collections: *class of persons*, and *ethnic group*. Similar results with respect to *class of persons* and *ethnic group* searches were observed in the main study, which analyzed collection searches in cultural heritage aggregation with U.S. history focus (Opening History), and in the pilot study, which targeted an aggregation of a much wider subject scope (IMLS Digital Collections and Content Collection Registry) that includes humanities, social sciences and sciences digital collections. This similarity allows to make a conclusion about generalizability of the *class of persons* and *ethnic group* search categories beyond the scholarly historians to a wider community of aggregation users. Unlike *collection*, these two search categories do not obviously fit under any of the higher-level entities in the FRBR model.

These results suggest that a model of collection-level subject access can be developed as a specific collection-level application of the FRBR model. In addition to FRBR subject entities (*work*, *person*, *corporate body*, *concept*, *object*, *event*, and *place*), the model of collection-level subject access can include *collection*, *class of persons*, and *ethnic group* subject entities. For such a model to further align with the context of use in aggregations of digital collections, it can include an agent (*user of aggregation*) and a "searches for" relation meaning the search term or attribute in the search query (Figure 30).

6.1.5 Theoretical Contributions

This study makes several theoretical contributions to the area of collection-level subject access. It is one of the first studies that collected empirical data -- both qualitative and quantitative -- about the collection-level information seeking behavior. This research tested the FRBR model of bibliographic universe in the context of collection-level searching in

aggregations of digital collections and resulted in suggesting a model of collection-level subject access (Figure 30). The study also developed and tested a definition of collection-level subject metadata richness as expression of digital collection's subject matter through mutually complementary values encoded in a variety of collection metadata fields and representing a variety of collection characteristics in the free-text *Description* fields.

6.1.6 Practical Implications for Development of Collection Metadata

A high level of engagement with collection metadata records demonstrated by this study as well as a high value placed at collection metadata by participants of this study supplies evidence for the added value of developing collection metadata records in aggregations of digital collections. Structured display of collection metadata records in their entirety to the end-users is found to be beneficial for the subject access, and this benefit should be considered by creator of aggregations.

This study has shown that the free-text *Description* fields and four structured subject-specific collection metadata fields (*Subjects, Geographic Coverage, Temporal Coverage,* and *Objects*) are crucial in providing subject access in aggregations of digital collections. These collection metadata fields should be applied more consistently in collection metadata in aggregations to ensure higher collection search retrieval results and to support browse functionality.

More formal recommendations or best practice guidance for creating rich free-text *Description* fields to describe digital collections in aggregations should be developed, for example as part of the *Framework of Guidance for Building Good Digital Collections* NISO standard (NISO, 2007). In addition to applicable item-level guidelines (e.g., *Cataloging Cultural*

Objects, 2008; *OSU Knowledge Bank Metadata Application Profile*, 2006; *Online Audiovisual Catalogers' Cataloging Policy*, 2002; *Dublin Core Usage Guide*[57]) and available documented collection-level practices (e.g., National Union Catalog of Manuscript Collections online datasheet for participating collections[58]) the findings of this study with respect to the emerging best practices in collection characteristics encoded in *Description* fields could be instrumental in developing these recommendations.

The findings of this study indicate great importance of applying a variety of subject-bearing collection metadata fields in describing collections to facilitate subject access in aggregations of digital collections. Even if the fields other than free-text *Description* are not displayed to the user, the value of the richness of "behind-the-scenes" subject metadata should not be underestimated.

6.2 Limitations and Future Research

This dissertation research makes contributions to the understanding of collection-level subject access and the role played in it by collection metadata, as well as to the testing of FRBR model of bibliographic universe against the real subject searching data in the context of aggregations of digital collections. However, the generalizability of some of this study's results is somewhat limited primarily due to its qualitative exploratory nature. Wherever possible, these limitations were minimized through triangulation of research methods and purposeful sampling. The limitations that were not alleviated in this study will be addressed in future research.

[57] http://dublincore.org/documents/2001/04/12/usageguide/sectb.shtml#description.
[58] http://www.loc.gov/coll/nucmc/lcforms.html

6.2.1 Limitations

The limitation of one of the major methods used in this dissertation research — content analysis — is the difficulty of achieving high consistency among coders due to the lack of agreement among people on the interpretation of categories. Intercoder reliability in this study was increased through development of detailed coding manual. The transaction log analysis method provides objective data on information system user actions but not on motivations and reasoning behind the actions. Reliability of transaction log analysis data is also impeded by the fact that transaction logs do not provide a way to separate the searches made by two very different user communities — end-users, such as scholarly historians, and digital resource developers. On the other hand, the major limitation of interviews is that they provide subjective information on perceived user experiences but lack objective data on actual experiences. Introducing an observation component into the interview helps to gain the more objective data on actual user experiences. Both interview and observation are obtrusive methods of data collection, and the results might be somewhat distorted because of this. When these methods are used together, like in this study, with unobtrusive methods like transaction log analysis and content analysis, however, the limitations of each of the approaches are minimized through triangulation.

Although I have been striving to get representation of the actual scholars, the interview/observation data has been used mainly to complement a comparative content analysis of collection-level metadata and transaction log analysis of user queries, and to fill out the real-life picture of the application and use of collection-level metadata. The sample for interview and observation component of the study was small but the direct match of the participants' research and teaching interests with the content of the cultural heritage aggregations studied and their knowledge of aggregations like American Memory ensured richness of collected data.

The study is also somewhat limited in its choice of aggregations of digital collections, as hundreds of other aggregations of different scope and scale exist in United States and abroad beyond the three selected for this study.[59]

6.2.2 Future Research

This dissertation study examined two important components of collection-level subject access in aggregations of digital collections: collection metadata that describes digital collections, and the users of these digital collection, or more specifically, user's interactions with an aggregation and collection metadata in the process of subject-based information discovery. There is a third important component in collection-level subject access — the agents who create collection metadata — that has been out of the scope of this study. Further research into each of these three components will extend understanding of collection-level subject access.

To make the results of comparative content analysis of collection metadata more generalizable, the sample of large-scale aggregations analyzed in this study should be expanded in future studies by including more aggregations of different scope — national (e.g., Memory of the Netherlands), state (e.g., Missouri Digital Heritage), and regional (e.g., Documenting the American South) — both in United States and abroad. In addition to aggregations of cultural heritage digital collections, the aggregations of other types of content (e.g., National Science Digital Library) also need to be analyzed. This study would answer such research questions as: How does the user's collection-level information seeking behavior differ in aggregations of

[59] See for example this extensive list of digital content aggregations http://oedb.org/library/features/250-plus-killer-digital-libraries-and-archives

different scope and focus? What are the differences between collection-level information seeking behavior of scientists, social scientists, and humanities scholars?

Studying information seeking behavior in aggregations applying the combination of content analysis with transaction log analysis is a promising approach for further research. Analyzing session-level user interactions in addition to query-level user interactions would add value to this research. Query-level analysis will provide important contextual information for individual interactions (e.g., subject browse, collection-level search) with the aggregations: search tactics, sequence of moves, query reformulation strategies, numbers of pages viewed in the session etc. More comprehensive comparison between collection-level and item-level user interactions would be beneficial for expanding the understanding of user interactions with aggregations. For example, this study focused on comparing collection-level search patterns to the patterns of item-level search in aggregation, while the detailed analysis of item browse interactions was not undertaken to compare it with collection browse. Similarly to content analysis of collection metadata, the more generalizable results will be produced by a comparative study of transaction log data from several aggregations of different size, scope, and focus.

A survey or series of interviews with those responsible for creating collection-level metadata in various aggregations would help to extend understanding on how the decisions are made about important issues including which collection metadata schema to follow, how to customize them, which metadata fields to use, what kinds of information to encode in which fields, which controlled vocabularies to use, and which fields then to display to the end-users of aggregations.

6.3 Figures and Tables

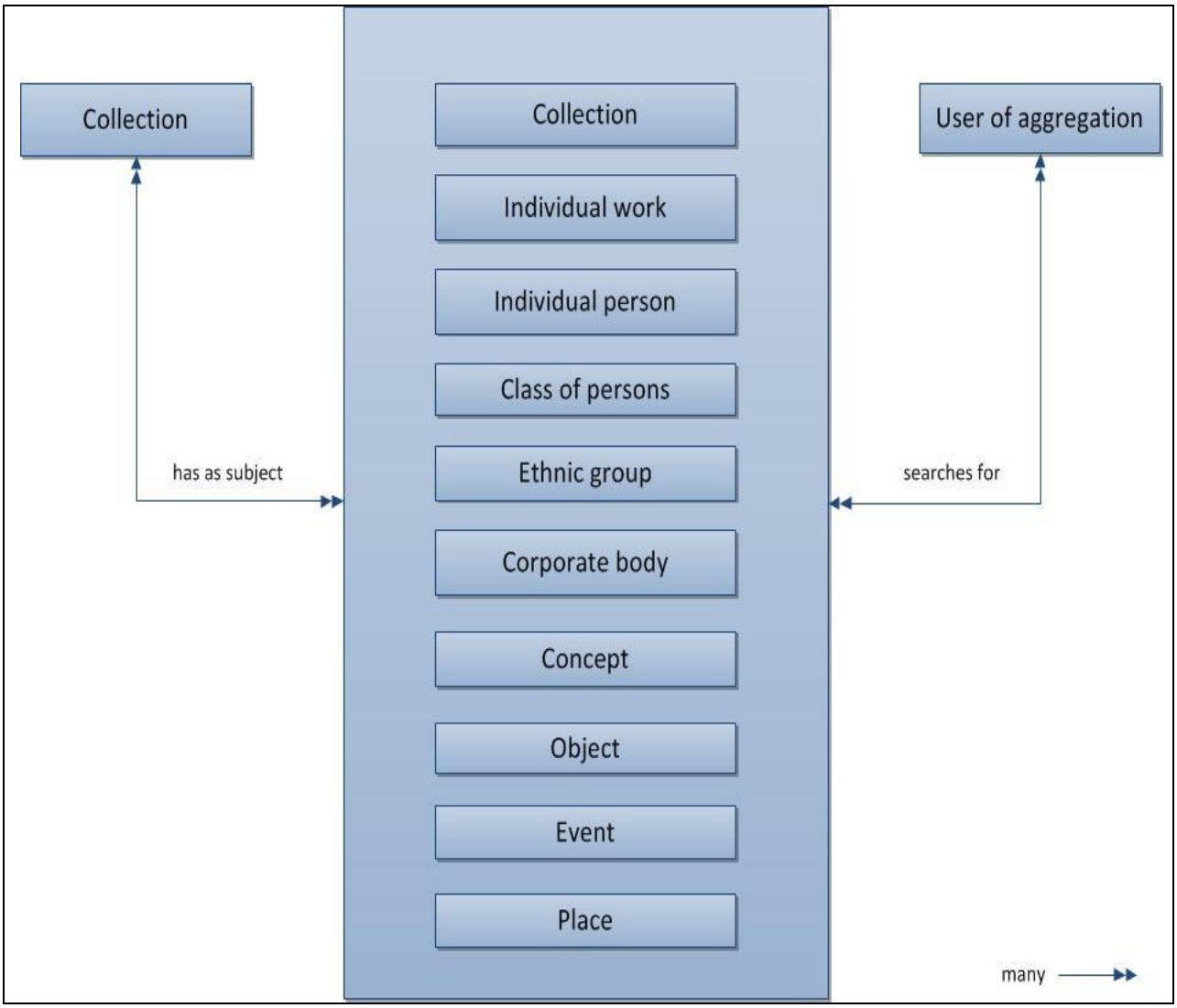


Figure 30. Model of collection-level subject access

Research Question(s)	Findings /Answers
How does collection-level metadata mediate scholarly subject access?	

• What is the variation in richness of collection-level subject metadata across collections and aggregations of digital collections?	• Between 5 and 7.8 collection properties (i.e., kinds of information about digital collection) encoded in the free-text *Description* field
	  • information about subjects, objects, geographic and temporal coverage is found the most often
	  • followed by collection development, creator of items, title, provenance, etc.
	• Mutually complementary information in free-text and controlled-vocabulary collection metadata fields

• How do scholarly users of cultural heritage aggregations approach collection-level information discovery?	• Browse more than search • Subject and geographic browse occur most often • Conduct search and browse at collection level less often than at item level • Search and browse, follow links, navigate back and force between collection and item records • Search more for research purposes, browse more for teaching purposes • Search mostly for objects, concepts, and places • Collection-level search terms somewhat differ from item-level search terms

• Which collection-level metadata fields provide scholarly users with the most valuable information to meet their needs?	• Scholars expect to see:
	  • Description
	  • Subjects
	  • Objects
	  • Geographic Coverage
	  • Temporal Coverage
	  • Provenance
	  • Size
	• Fields that provide most matches to user search terms:
	  • **Description**
	  • **Subjects**
	  • Objects
	  • Geographic Coverage
	  • Temporal Coverage

• How does collection-level user search data fit the FRBR model?	• FRBR Group 3 of entities (*concept*, *object*, and *place*) match most of the collection-level search categories • No *family* searches (FRBR Group 2 of entities) observed • Impossible to distinguish between *work* and other Group 1 entities (*expression*, *manifestation*, *item*) in collection-level search categories • Users search for item-level or collection-level *work* • Users search for *ethnic group* and *class of persons* — two categories not covered by FRBR model

Table 16. Research questions and findings/answers of this study

Bibliography

Agosti, M. et al. (2007). Analysing HTTP logs of a European DL Initiative to maximize usage and usability. In D.H.-L. Goh et al. (Eds.), *ICADL 2007, LNCS 4822*. Berlin: Springer-Verlag, pp. 35-44.

Allen, B. (1991). Cognitive research in information science: implications for design. *Annual Review of Information Science and Technology, 26*, 3-37.

Allen, B., & Reser, D. (1990). Content analysis in library and information science research. *Library & Information Science Research, 12*, 251-262.

Allen, B., & Sutton, B. (1993). Exploring the intellectual organization of an interdisciplinary research institute. *College & Research Libraries, 54*, 499-515.

Atherton, P. (1978). *Books Are for Use: Final Report of the Subject Access Project to the Council on Library Resources*. Syracuse, NY: School of Information Studies.

Bates, M. (1972). *Factors affecting subject catalog search success*. Ph.D. dissertation. University of California, Berkeley.

Bates, M. (1989). Rethinking subject cataloging in the online environment. *Library Resources & Technical Services, 33*(4), 400.

Bates, M. (1996). The Getty end-user online searching project in the humanities: Report no. 6: Overview and conclusions. *College & Research Libraries, 57*, 514—523

Bawden, D., & Vilar, P. (2006). Digital libraries: to meet or manage user expectations. *Aslib Proceedings, 58*(4), 346-354.

Beall, J. (2006). *Metadata Schemes Points of Comparison*. Retrieved October 22, 2008, from http://eprints.rclis.org/archive/00005544/01/comparingschemes.pdf.

Becker, N. (2003). Google in perspective: understanding and enhancing student search skills. *New Review of Academic Librarianship, 9*, 84-100.

Beitzel, S., Jensen, E., Chowdhury, A., Frieder, O., & Grossman, D. (2007). Temporal analysis of a very large topically categorized Web query log. *Journal of the American Society for Information Science and Technology, 58*(2), 166-178.

Borgman, C. (1986). Why are online catalogs hard to use? Lessons learned from information retrieval studies. *Journal of American Society for Information Science, 37* (6), 387-400.

Borgman, C. (1996). Why are online catalogs still hard to use? *Journal of American Society for Information Science, 47* (7), 493-503.

Brophy, J., & Bawden, D. (2005). Is Google enough? Comparison of an Internet search engine with academic library resources. *Aslib Proceedings, 57*(6), 498-512.

Buchanan, G., Cunningham, S. J., Blandford, A., Rimmer, J., & Warwick, C. (2005). Information seeking by humanities scholars. In *Proceedings of the European Conference on Digital Libraries (ECDL2005)*, 218-229.

Buckland, M. (1999). Vocabulary as a central concept in library and information science. In T. Arpanac et al. (Eds.), *Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities.* Retrieved October 1, 2007, from http://www.sims.berkeley.edu/~buckland/colisvoc.htm.

Case, D.O. (1991). The collections and use of information by some American historians: A study of motives and methods. *Library Quarterly, 61*, 61-82.

Chan, L., & Hodges, T. (2000). Entering the millennium: A new century for LCSH. *Cataloging and Classification Quarterly, 29* (1-2), 225-234.

Cleveland, G. (1998). *Digital Libraries: Definitions, Issues and Challenges. UDT Occasional Paper, 8.* Retrieved September 27, 2007, from http://www.ifla.org/VI/5/op/udtop8/udtop8.htm.

Cochrane, P. (1979). Universal Subject Access (USA): can anyone do it? In *Redesign of Catalogs and Indexes for Improved Online Subject Access: selected papers of Pauline A. Cochrane*, Phoenix, Ariz.: Oryx Press, 1985, pp. 223-238.

Cochrane, P. (1986). *Improving LCSH for Use in Online Catalogs*. Colorado Springs, CO: Libraries Unlimited.

Cochrane, P. (2000). Improving LCSH for use in online catalogs revisited: What progress has been made? What issues still remain? *Cataloging and Classification Quarterly, 29* (1/2), 73-89.

Cole, T., & Shreeves, S. (2004). Search and discovery across collections: The IMLS Digital Collections and Content project. *Library Hi Tech, 22*(3), 307-322.

Connaway, L., Johnson, D., & Searing, S. (1997). Online catalogs from the user's perspective: the use of focus group interviews. *College and Research Libraries, 58* (September), 403-420.

Covey, D. (2002). Usage studies of electronic resources. In *Usage and Usability Assessment: Library Practices and Concerns*. Washington, DC: Digital Library Federation and Council on Library and Information Resources. (CLIR Report 105.) Retrieved May 20, 2008, from http://www.clir.org/PUBS/reports/pub105/section3.html

Currall, J., Moss, M., & Stuart, S. (2004). What is a collection? *Archivaria*, (58), 131-146.

Curl, M. (1995). Enhancing subject and keyword access to periodical abstracts and indexes: Possibilities and problems. *Cataloging & Classification Quarterly, 20*(4), 45-55.

Cutter, C. (1904). *Rules for a Dictionary Catalog*. (4th Ed.). Washington, DC: Govt. Printing Office.

Delsey, T. (2005). Modeling subject access: Extending the FRBR and FRANAR conceptual models. *Cataloging and Classification Quarterly, 39* (3/4), 49-61.

Dervin, B. (1983). *An Overview of Sense-making Research: Concepts, Methods and Results to Date.* Paper presented at the International Communications Association Annual Meeting. Dallas, Texas.

Drabenstott, K. (1991). Online Catalog User Needs and Behaviors. In *Think Tank on the Present and Future of the Online Catalog: Proceedings*, edited by Noelle Van Pulis, 59-84. Chicago: Reference and Adult Services Division, American Library Association.

Drabenstott, K., & Weller, M. (1996). Failure analysis of subject searches in a test of a new design for subject access to online catalogs. *Journal of American Society for Information Science, 47* (7), 519-537.

Duff, W.M., & Johnson, C.A. (2002). Accidentally found on purpose: Information-seeking behaviors of historians in archives. *Library Quarterly, 72*(4), 472-496.

Duval, Erik et al. (2002). Metadata Principles and Practicalities. *D-Lib Magazine, 8(4)*. Retrieved October 21, 2008, from http://www.dlib.org/dlib/april02/weibel/04weibel.html.

Ellis, D., & Oldman, H. (2005). The English literature researcher in the age of the Internet. *Journal of Information Science, 31* (1), 29-36.

ePrints DCMI Application Profile and Cataloguing Guidelines (2007). Retrieved April 1, 2007, from http://www.ukoln.ac.uk/repositories/digirep/index/EPrints_Application_Profile.

Farradine, J. (1970). Analysis and organization of knowledge for retrieval. *Aslib Proceedings, 22*(12), 607-616.

Fast, K., & Campbell, D. (2004). I still like Google: university student perceptions of searching OPACs and the Web. In *Proceedings of the 67th ASIS&T Annual Meeting* (Providence, R.I.: American Society for Information Science and Technology)

Fidel, R. (1988). Factors affecting the selection of search keys. In *Proceedings of the 51st annual meeting of the American Society for Information Science, volume 25, Atlanta, Georgia, 23-27 October 1988* Medford, NJ: Learned Information, pp. 76-79.

Fidel, R. (1992). Who needs controlled vocabulary? *Special libraries, 83*(1), 1-9.

Foulonneau, M., Cole, T., Habing, T., & Shreeves, S. (2005). Using collection descriptions to enhance an aggregation of harvested item-level metadata. In *JCDL 05: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries,* ACM Press, New York, NY, USA, pp. 32-41.

Garrett, J. (2007). Subject headings in full-text environments: the ECCO experiment. *College & Research Libraries, 68*(1), 69-81.

Gault, L., Shultz, M., & Davies, K. (2002). Variations in medical subject headings (MeSH) mapping: From the natural language of patron terms to the controlled vocabulary of mapped lists. *Journal of the Medical Library Association (JMLA), 90*(2), 173-180.

Geisler, G., Giersch, S., McArthur, D., and McCelland, M. (2002). Creating virtual collections in digital libraries: benefits and implementation issues. *Joint Conference on Digital Libraries 2002*, Portland, OR, pp. 210-218. Retrieved May 20, 2010, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.116.5329&rep=rep1&type=pdf.

Godby, C. J., Smith, D. & Childress, E. (2003). Two Paths to Interoperable Metadata. *DC-2003: Supporting Communities of Discourse and Practice-Metadata Research & Applications*. Retrieved October 10, 2008, from http://www.oclc.org/research/publications/archive/2003/godby-dc2003.pdf.

Greenberg, J. (2001). Automatic query expansion via lexical-semantic relationships. *Journal of the American Society for Information Science, 52*(5), 402-415.

Greenberg, J. (2003). Metadata and the World Wide Web. *Encyclopedia of Library and Information Science* 1876-1888. Retrieved October 21, 2008, from http://www.informaworld.com.proxy2.library.uiuc.edu/10.1081/E-ELIS-120008663.

Griffith, J., & Brophy, P. (2005). Student searching behavior and the Web: Use of academic resources and Google. *Library Trends, 53*(4), 539-554.

Gross, T., & Taylor, R. (2005). What have we got to lose? The effect of controlled vocabulary on keyword searching results. *College and Research Libraries, 66*(3), 212-230.

Harum, S. (2008). Personal conversation.

Heaney, M. (2000). *An Analytical Model of Collections and Their Catalogues*. Retrieved January 25, 2008, from http://www.ukoln.ac.uk/metadata/rslp/model/amcc-v31.pdf.

Heery, R., & Patel, M. (2000). Application profiles: mixing and matching metadata schemas. *Ariadne, 25*. Retrieved October 28, 2008, from http://www.ariadne.ac.uk/issue25/app-profiles/intro.html.

Henri, F. (1992). Computer conferencing and content analysis. *Collaborative Learning through Computer Conferencing: The Najaden Papers*. A. R. Kaye. New York, Springer**,** 115-136.

Hembrooke, H., Granka, L., Gay, G., & Liddy, E. (2005). The effects of expertise and feedback on search term selection and subsequent learning. *Journal of the American Society for Information Science and Technology, 56*(8), 861-871.

Hildreth, C. (1989). *Intelligent Interfaces and Retrieval Methods for Subject Searching in Bibliographic Systems*, Prepared for the Library of Congress. Washington, DC: Cataloging Distribution Service.

Hildreth, C. (1995). *Online Catalog Design Models: Are We Moving in the Right Direction?* Retrieved May 20, 2010 from http://myweb.cwpost.liu.edu/childret/clrintro.html.

Hildreth, C. (1997). The use and understanding of keyword searching in a university online catalog. *Information Technology and Libraries, 16*(2), 52-62.

Hill, L. et al. (1999). Collection metadata solutions for digital library applications. *Journal of the American Society for Information Science, 50*(13), 1169-1181.

Hjørland, B. (1997). The concept of subject or subject matter and basic epistemological positions. In *Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science*. Westport CT: Greenwood Press, 55-103.

Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science, 44*(3), 161-174.

Hutt, A., & Riley, J. (2005). Semantics and Syntax of Dublin Core Usage in Open Archives Initiative Data Providers of Cultural Heritage Materials. In *Fifth ACM/IEEE-CS Joint Conference on Digital Libraries 2005* (262-270). New York: ACM Press.

IFLA Study Group on the Functional Requirements for Bibliographic Records (1998). *Functional Requirements for Bibliographic Records: Final report.* Munchen: K.G.Saur. Retrieved April 1, 2007, from http://www.ifla.org/VII/s13/frbr/frbr.pdf.

IFLA Study Group on the Functional Requirements for Bibliographic Records (2008). *Functional Requirements for Bibliographic Records: Final report: As amended and corrected through February 2008*. Retrieved March 17, 2008, from http://www.ifla.org/VII/s13/frbr/frbr2008.pdf.

IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR) (2007). *Functional Requirements for Authority Data: A Conceptual Model*. 2nd draft. Retrieved April 12, 2007, from http://www.ifla.org/VII/d4/FRANAR-ConceptualModel-2ndReview.pdf.

Information Behaviour of the Researcher of the Future (2008). Retrieved April 18, 2008, from http://www.jisc.ac.uk/media/documents/programmes/reppres/gg_final_keynote_11012008.pdf.

Ingwersen, P. (1994). Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In *Proceedings of the SIGIR 1994,* 101-110.

Institute for Museum and Library Services (2003). *Assessment of End-User Needs in IMLS-funded Digitization Projects*. Retrieved March 31, 2008 from www.imls.gov/pdf/userneedsassessment.pdf.

International Council of Museums/CIDOC (2007). *Definition of the CIDOC Conceptual Reference Model*: version 4.2.2. Retrieved January 25, 2008 from http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.2.pdf.

Jackson, A. S., Han, M. J., Groetsch, K., Mustafoff, M., & Cole, T. W. (2008). Dublin Core metadata harvested through OAI-PMH. *Journal of Library Metadata, 8 (1).*

Jackson, S. (1958). *Catalog Use Study: Director's Report*. Chicago: American Library Association.

Jansen, B. (2006). Search log analysis: What is it, what's been done, how to do it. *Library and Information Science Research,* 28(3), 407-432. Retrieved October 22, 2008, from http://ist.psu.edu/faculty_pages/jjansen/academic/pubs/jansen_search_log_analysis.pdf.

Jansen, B. (2008). The methodology of search log analysis. In Bernard J. Jansen, Amanda Spink, & Isak Taksa (Eds.), *Handbook of Research on Web Log Analysis*, Hershey, PA: Information Science Reference, pp. 100-123. Jansen, B., & Pooch, U. (2001). A review of Web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology, 52*(3), 235-246.

Jansen, B., & Pooch, U. (2001). A review of Web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology, 52*(3), 235-246.

Jansen, B., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management, 42*(1), 248-263.

Jansen, B., Spink, A., & Koshman, S. (2007). Web searcher interaction with the Dogpile.com metasearch engine. *Journal of the American Society for Information Science and Technology, 58*(5), 744-755.

Jansen, B., Spink, A., & Pedersen, J. (2004). The effect of specialized multimedia collections on Web searching. *Journal of Web Engineering, 3* (3/4), 182-199.

Johnston, P. (2003). *Report from Meeting of DC CD WG at DC-2003*. Retrieved October 28, 2008, from http://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind0310&L=DC-COLLECTIONS&D=0&I=-3&P=59.

Johnston, P., & Robinson, B. (2002). Collections and collection description. *Collection Description Focus Briefing Paper, 1*. Retrieved June 20, 2006 from http://www.ukoln.ac.uk/cd-focus/briefings/bp1/bp1.pdf.

Kipp, M. (2006) Complementary or Discrete Contexts in on-line indexing: A comparison of user, creator and intermediary keywords. *Canadian Journal of Information and Library Science*. Retrieved March 28, 2008 from http://dlist.sir.arizona.edu/1533.

Krikelas, J. (1972). Catalog use studies and their implications. *Advances in Librarianship, 3,* 195-220.

Lagoze, C., & Fielding, D. (1998). Defining collections in distributed digital libraries. *D-Lib Magazine,* (November). Retrieved June 20, 2006 from http://webdoc.gwdg.de/edoc/aw/d-lib/dlib/november98/lagoze/11lagoze.html.

Larson, R. (1991a). Between Scylla and Charybdis: Subject searching in online catalogs. *Advances in Librarianship, 15,* 175-236.

Larson, R. (1991b). The decline of subject searching: long-term trends and patterns of index use in an online catalog. *Journal of American Society for Information Science, 42*(3), 197-215.

Larson, R. (1991c). Classification clustering, probabilistic information retrieval, and the online catalog. *Library Quarterly, 61*(2), 133-173.

Lee, H. (2000). What is a collection? *Journal of the American Society for Information Science, 51*(12), 1106-1113.

Lee, H. (2003). Information spaces and collections: Implications for organization. *Library & Information Science Research, 25*(4), 419-436.

Lee, H. (2005). The concept of collection from the user's perspective. *Library Quarterly, 75*(1), 67-85.

Lee, J., Renear, A., & Smith, L. (2006). Known-item searching: Variations on a concept. *Proceedings of the 69th ASIS&T Annual Meeting*, 3-8 November 2006, Austin, Texas.

Lipetz, B. (1970). *User Requirements in Identifying Desired Works in a Large Library.* New Haven, CT: Yale University Library.

Lubetzky, S. (1960). *Code of Cataloging Rules: Author and Title Entry: An Unfinished Draft*. Chicago, IL: American Library Association.

Lynch, C. (2002). Digital collections, digital libraries, and the digitization of cultural heritage information. *First Monday, 7*(5).

Lynch, C., & Garcia-Molina, H. (1995). *Interoperability, Scaling, and the Digital Libraries Research Agenda: a Report on the May 18-19, 1995 IITA Digital Libraries Workshop*. Retrieved September 25, 2007 from http://www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html.

Macgregor, G. (2003). Collection-level descriptions: metadata of the future? *Library Review, 52*(6), 247-250.

Manoff, M. (2000). Hybridity, mutability, multiplicity: Theorizing electronic library collections. *Library Trends, 48*(4), 857-876.

Marchionini, G., Dwiggins, S., Katz, A., & Lin, X. (1993). Information seeking in full-text end-user-oriented search systems: the roles of domain and search expertise. *Library and Information Science Research, 15*(1), 35-69.

Markey, K. (1984). *Subject Searching in Library Catalogs*. Dublin, Ohio: OCLC.

Markey, K. (2007). Users and uses of bibliographic data: presentation. *Library of Congress Working Group on the Future of Bibliographic Control Meeting*, March 8, 2007, Mountain View, California.

Markey, K., & Calhoun, K. (1987). Unique words contributed by MARC records with summary and/or contents notes. In *ASIS '87, Proceedings of the 50th ASIS Annual Meeting*, edited by Ching-chih Chen, 153-162. Medford, NJ: Learned Information.

Markey, K., & Demeyer, A. (1986). *Dewey Decimal Online Classification Project*. Dublin, OH: OCLC.

Matthews, J., Lawrence, G., & Ferguson, D. (Eds.), (1983). *Using Online Catalogs: A Nationwide Survey*: *A Report of a Study Sponsored by the Council on Library Resources.* New York, NY: Neal-Schumann.

Maxwell, R. L. (2008). *FRBR: A Guide for the Perplexed*. Chicago, IL: American Library Association.

Miller P. (2000). Collected wisdom: some cross-domain issues of collection-level description. *D-Lib Magazine, 6*(Sept.) Retrieved November 25, 2007 from http://www.dlib.org/dlib/september00/miller/09miller.html.

Moen, W.E., & Benardino, P. (2003). Assessing metadata utilization: an analysis of MARC content designation use. In *Proceedings of the International Conference on Dublin Core and Metadata Applications* (Seattle, WA, Sept. 28 - Oct.2, 2003). Retrieved October 22, 2008, from http://dc2003.ischool.washington.edu/Archive-03/03moen.pdf.

Muddamalle, M. (1998). Natural language versus controlled vocabulary in information retrieval: A case study in soil mechanics. *Journal of American Society for Information Science, 49*(10), 881-887.

NISO Framework Advisory Group (2007). *A Framework of Guidance for Building Good Digital Collections*. 3rd edition. Bethesda, MD: National Information Standards Organization. Retrieved June 16, 2010, from http://www.niso.org/publications/rp/framework3.pdf.

Nowick, E., & Mering, M. (2003). Comparisons between Internet users' free-text queries and controlled vocabularies: a case study in water quality. *Technical Services Quarterly, 21*(2), 15-32.

Palmer, C., Knutson, E., Twidale, M., & Zavalina, O. (2006). Collection definition in federated digital resource development. *Proceedings of the 69th ASIS&T Annual Meeting*, 3-8 November 2006, Austin, Texas.

Palmer, C., Zavalina, O., & Mustafoff, M. (2007). Trends in metadata practices: a longitudinal study of collection federation metadata. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (Vancouver, Canada, June 19-23, 2007), 386-395.

Palmer, R. (1970). *User Requirements of a University Library Card Catalog*. Unpublished dissertation, University of Michigan.

Panizzi, A. (1841). Rules for the compilation of the catalogue. *Catalog of Printed Books in the British Museum, 1*, [v]-ix.

Park, J. (2005). Semantic interoperability across digital image collections: A pilot study on metadata mapping. In CAIS/*ACSI 2005 Data, Information, and Knowledge in a Networked World*, edited by Liwen Vaughan. Proceedings of the 2005 annual conference of the Canadian Association for Information Science held with the Congress of the Social Sciences and Humanities of Canada at the University of Western Ontario, London, Ontario, June 2-4, 2005.

Pennanen, M., Serola, S., & Vakkari, P. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing and Management, 39*(3), 445-463.

Peters, T. (1991). *Online Catalog: a Critical Examination of Public Use*. Jefferson, NC: McFarland.

Peters, T. (1993). The history and development of transaction log analysis. *Library Hi Tech, 11*(2), 41-66.

Raymond, M. (2008). *My Friend Flickr: A Match Made in Photo Heaven.* http://www.loc.gov/blog/?p=233.

Reynolds, J. (1995). A brave new world: user studies in the humanities enter the electronic age. *Reference Librarian, 49*(50), 61-81.

Riesthuis, G., & Žumer, M. (2004). FRBR and FRANAR: subject access. *8th International ISKO Conference.* Retrieved October 4, 2007, from http://www.ucl.ac.uk/isko2004/sysweb/4bRiesthuisZumer.ppt.

Shreeves, S.L., Riley, J., & Milewicz, L. (2006). Moving towards sharable metadata. *First Monday, 11* (8) Retrieved October 4, 2008, from http://firstmonday.org/issues/issue11_8/shreeves/index.html.

Slone, D. (2000). Encounters with the OPAC: On-line searching in public libraries. *Journal of the American Society for Information Science, 51*(8), 757-773.

Spink, A., & Jansen, B. (2004). A Study of Web Search Trends. *Webology, 1*(2). Retrieved May 17, 2010, from http://webology.ir/2004/v1n2/a4.html.

Spink, A., Jansen, B., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *IEEE Computer, 35*(3), 107-111.

Spink, A., Wolfram, D., Jansen, B., & Saracevic, T. (2001). Searching the web: the public and their queries. *Journal of the American Society for Information Science and Technology, 52*(3), 226-234.

Stone, S. (1982). Humanities scholars: information needs and uses. *Journal of Documentation, 38,* 292-313.

Tagliacozzo, R., & Kochen, M. (1970). Information-seeking behavior of catalog users. *Information Storage and Retrieval, 6,* 363-381.

Taube, M. (1953). *Studies in Coordinate Indexing.* Washington D.C.: Documentation Incorporated.

Tibbo, H. (2003). Primarily history in America: How U.S. historians search for primary materials at the dawn of the digital age. *The American Archivist, 66* (1), 9-50.

Waters, D. (1998). What are digital libraries? *CLIR Issues, July/August*. Retrieved September 25, 2007 from http://www.clir.org/pubs/issues/issues04.HTML.

Weare, C., & Lin, W.-Y. (2000). Content analysis of the World Wide Web: opportunities and challenges. *Social Science Computer Review, 18* (3), 272-292.

Weinberg, B. (1995). Why postcoordination fails the researcher. *The Indexer, 19,* 155-159.

Wildemuth, B. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology, 55*(3), 246-258.

Wormell, I. (1981). Subject Access Project: the use of book indexes for subject retrieval system in libraries. *International Forum on Information and Documentation, 6*(4), 24-28.

Zavalina, O. (2007). Collection-level user searches in federated digital resource environment. In *Proceedings of the 70th ASIS&T Annual Meeting* (Milwaukee WI, Oct. 19-24, 2007).

Zavalina, O., Palmer, C.L., Jackson, A.S., & Han, M.-J. (2008a2008a). Assessing descriptive substance in free-text collection-level metadata. In *Proceedings of the 8th International Conference on Dublin Core and Metadata Applications* (Berlin, Germany, Sept. 22-26, 2008).

Zavalina, O., Palmer, C.L., Jackson, A.S., & Han, M.-J. (2008b). Evaluating descriptive richness in collection-level metadata. *Journal of Library Metadata, 8*(4), 263-292.

Zeng, M., & Salaba, A. (2005). Toward an international sharing and use of subject authority data. *FRBR Workshop*, OCLC, 2005. Retrieved October 4, 2007, from http://www.oclc.org/research/events/frbr-workshop/presentations/zeng/Zeng_Salaba.ppt.

Zhang, X., Anghelescu, H., & Yuan, X. (2005). Domain knowledge, search behavior, and search effectiveness of engineering and science students: An exploratory study. *Information Research, 10*(2).

Zhang, Y., & Salaba, A. (2007a). Critical issues and challenges facing FRBR research and practice. *Bulletin of the American Society for Information Science and Technology, Aug./Sept.*

Zhang, Y., & Salaba, A. (2007b). User research and testing of FRBR prototype systems: Poster. *70th ASIS&T Annual Meeting* (Milwaukee WI, Oct. 19-24, 2007).

Appendix A. Interview/Observation Guide

Section # 1

1. Have you ever used the Opening History (OH) aggregation of digital collections for information seeking related to your recent major research project (e.g., book, chapter, research paper, dissertation proposal/thesis, etc.)?
    a. If yes, please answer questions in Section 2. PLEASE USE THE OPENING HISTORY SITE TO HELP YOU RECALL AND SHOW ME WHAT YOU DID.
    b. If no, let us skip to Section 4.
2. Have you ever used the American Memory aggregation of digital collections for information seeking related to your recent major research project?
    a. If yes, please answer questions in Section 3. PLEASE USE THE AMERICAN MEMORY SITE TO HELP YOU RECALL AND SHOW ME WHAT YOU DID.
    b. If no, let us skip to Section 4.

Section # 2. What was your experience searching/browsing Opening History (OH)?

1.Which research project/task have you used OH for? When did your interaction with OH occur? Was there a single or multiple interactions in the course of working on that project/task?

2.What type of information did you need to find?

3.What was the subject area of your research/task? Were you working within or outside your regular research area?

4.What was your purpose in using OH: to find any relevant information on a subject, to locate a specific item/collection that you already knew about, other?

5.Did you start with search or browse?

6.If you used browse, which browse mode(s) did you select (e.g., subject, object, place, collection title, institution) and why? Was there any difference between your usual online browse approach (e.g., when searching in *Google, Historical Abstracts*, etc.) and the one(s) you used in OH? How would you rate usefulness of the option to browse by subject in locating needed information in OH aggregation?

7.If you used search, which search mode(s) did you use (e.g., simple, advanced, "collection only" search)? Was there any difference between your usual online search mode selection (e.g., when searching in *Google, Historical Abstracts*, etc.) and the one(s) you used searching in OH?

8.How did you formulate search queries?

9.If you used "collection only" search, did you find it useful? Has your search retrieved any results? If so, how did information found in the collection-level record influence your search

process for the item(s)? Which specific fields of the collection-level records you noticed that contained the match(es) to your search (e.g., *Description, Geographic Coverage, Time Period, Other Topics* etc.)?

10.If you used advanced search, which search approach(es) did you use (e.g., author, title, subject)?

11.When using simple or advanced search, have you looked at the collection-level record(s) for collection(s) containing the search hits? If so, how did information found in the collection-level record influence your search process for the item(s)? Which specific fields of the collection-level records you noticed that contained the match(es) to your search (e.g., *Description*, *Geographic Coverage, Time Period*, *Other Topics*, etc.)?

12.What are the major subject headings (from whatever thesaurus you use/ are familiar with) relevant to the topic of your research? The broader/narrower terms? On a scale of 10 (where 1 is not useful at all, and 10 is very useful), how would you rate usefulness of the subject headings in locating needed information in OH?

Section # 3. <u>What was your experience searching/browsing American Memory?</u>

1. Which research project/task have you used American Memory for? When did your interaction with American Memory occur? Was there a single or multiple interactions in the course of working on that project/task?
2. What type of information did you need to find?
3. What was the subject area of your research/task? Were you working within or outside your regular research area?
4. What was your purpose in using American Memory: to find any relevant information on a subject, to locate a specific item/collection that you already knew about, other?
5. Did you start with search or browse?
6. Which search approach did you use (e.g., "Search all collections", "Search selected collections")? Was there any difference between your usual online search approach (e.g., when searching in *Google*, *Historical Abstracts*, etc.) and the one(s) you used searching in American Memory? How do you think the two different options, if available — for collections search and for item search — benefit you experience?
7. How did you formulate search queries?
8. What are the major subject headings (from whatever thesaurus you use/ are familiar with) relevant to the topic of your research? The broader/narrower terms? On a scale of 10 (where 1 is not useful at all, and 10 is very useful), how would you rate usefulness of the subject headings in locating needed information in American Memory?
9. Which role did collections browsing play in your information seeking? Which browsing options did you select, if any (e.g., by title, by topic, by time period, by place, by item type) and why? How would you rate usefulness of the option to browse collections in locating needed information American Memory?

10. Have you looked at the collection-level record(s) for collection(s) found through search or browsing? If so, how did information found in the collection-level record influence your search process for the item(s)? Which specific fields of the collection-level records influence you search process for the item(s) (e.g., *About this collection*, *Building the digital collection*, *Rights and reproductions*, etc.)?
11. In your opinion, would your experience using American Memory be any different if American Memory had more structured/detailed/consistent collection-level descriptions? If yes, what would be the difference?

Section # 4. Let us observe your search and record your experiences in these two aggregations:

1. What type of digital cultural heritage collections/objects (e.g., collections containing photographs, letters, journal entries, books, government documents, oral histories, digitized physical objects like pottery or costume, etc.) would be important/useful to find for a major current or recent project of yours (e.g., book, chapter, research paper, dissertation proposal/thesis, etc.)?
2. What is the nature of research project/task?
3. What is the subject area of your research/task? Is it within or outside your regular research area?
4. Do you need to find any relevant information on a subject, to locate a specific item/collection that you already know about, other?
5. What are the specific terms that you would use for a search?
6. Let us explore OH aggregation http://imlsdcc.grainger.uiuc.edu/history/ using your preferred search/browse approach(es).
7. Let us explore American Memory aggregation http://memory.loc.gov/ammem/index.html using your preferred search/browse approach(es).
8. How would you characterize and compare your experience in two aggregations? (see questions 6-12 in Section 2, questions 6-11 in Section 3).

    Opening History

    a. If you used browse, which browse mode(s) did you select (e.g., subject, object, place, collection title, institution) and why? Was there any difference between your usual online browse approach (e.g., when searching in *Google*, *Historical Abstracts*, etc.) and the one(s) you used in OH? How would you rate usefulness of the option to browse by subject in locating needed information in OH aggregation?

    b. If you used search, which search mode(s) did you use (e.g., simple, advanced, "collection only" search)? Was there any difference between your usual online search mode selection (e.g., when searching in *Google*, *Historical Abstracts*, etc.) and the one(s) you used searching in OH?

    c. How did you formulate search queries?

d. If you used "collection only" search, did you find it useful? Has your search retrieved any results? If so, how did information found in the collection-level record influence your search process for the item(s)? Which specific fields of the collection-level records you noticed that contained the match(es) to your search (e.g., *Description, Geographic Coverage, Time Period, Other Topics* etc.)?

e. If you used advanced search, which search approach(es) did you use (e.g., author, title, subject)?

f. When using simple or advanced search, have you looked at the collection-level record(s) for collection(s) containing the search hits? If so, how did information found in the collection-level record influence your search process for the item(s)? Which specific fields of the collection-level records you noticed that contained the match(es) to your search (e.g., *Description, Geographic Coverage, Time Period, Other Topics*, etc.)?

g. What are the major subject headings (from whatever thesaurus you use/ are familiar with) relevant to the topic of your research? The broader/narrower terms? On a scale of 10 (where 1 is not useful at all, and 10 is very useful), how would you rate usefulness of the subject headings in locating needed information in OH?

American Memory

a. Which search approach did you use (e.g., "Search all collections", "Search selected collections")? Was there any difference between your usual online search approach (e.g., when searching in *Google*, *Historical Abstracts*, etc.) and the one(s) you used searching in American Memory? How do you think the two different options, if available — for collections search and for item search — benefit you experience?

b. How did you formulate search queries?

c. What are the major subject headings (from whatever thesaurus you use/ are familiar with) relevant to the topic of your research? The broader/narrower terms? On a scale of 10 (where 1 is not useful at all, and 10 is very useful), how would you rate usefulness of the subject headings in locating needed information in American Memory?

d. Which role did collections browsing play in your information seeking? Which browsing options did you select, if any (e.g., by title, by topic, by time period, by place, by item type) and why? How would you rate usefulness of the option to browse collections in locating needed information American Memory?

e. Have you looked at the collection-level record(s) for collection(s) found through search or browsing? If so, how did information found in the collection-level record influence your search process for the item(s)? Which specific fields of the collection-level records influence you search process for the item(s) (e.g., *About this collection, Building the digital collection, Rights and reproductions*, etc.)?

f. In your opinion, would your experience using American Memory be any different if American Memory had more structured/detailed/consistent collection-level

descriptions? If yes, what would be the difference? EXAMPLES: Ansel Adams collection records in OH and AM

9. Have you found what you were looking for? Which search/browsing tools and features were helpful, which did not help? Which role did collection-level records in OH and American Memory play in your information discovery?

Final (Optional) Question: Would you like to add something or comment on your impressions of the interview/observation?

Appendix B. Interview/Observation Informed Consent Form

My name is Oksana Zavalina. I am a doctoral student in Graduate School of Library and Information Science. You are invited to participate in my dissertation research conducted under supervision of Dr. Carole Palmer. The focus of my study is to learn how collection-level metadata represents digital collections in aggregations of digital content and how it serves the needs of the scholarly users. The experiences and insights of academic historians (in particular, faculty and doctoral students) doing research on aspects of US history will form the basis for this research.

If you decide to participate, you will be asked to discuss issues about your use of aggregations of digital collections in an interview and observation session. To help me understand better whether and how the collection-level descriptions in the Opening History (OH) http://imlsdcc.grainger.uiuc.edu/history and American Memory (AM) http://memory.loc.gov/ammem/index.html aggregations are helpful in finding digital objects used in history research, I would like to audiotape you while you are using and commenting on the OH and AM systems. For audio-recording I will use the Camtasia software, which also keeps a record of the interaction (what you click on, etc.) so that I can learn more about the problems with Opening History's current design. I will also take notes and ask you what you think about how the OH system could be improved. Interview and observation together will last up to 1 hour.

Your participation in this research project is entirely voluntary, with no risks besides those of everyday life. You may not benefit from participation, but your participation will benefit the general knowledge about digital libraries and their value for scholarly research, particularly history research. The results of this study will be disseminated as my PhD thesis and conference presentations. I will not use your name or any identifiable information in any research reports or presentations, unless you ask me specifically to mention your real name. I will keep the recordings of our interview and observation secure until the project is finished, then I will destroy the recordings.

Your decision whether or not to participate will not affect your future relations with the University of Illinois at Urbana-Champaign. You are under no obligation to participate in the study. You are free to (a) discontinue participation in the study at any time, (b) request that the audio recorder be turned off at any time, and (c) pass on any question you do not want to answer. If you decide not to participate in this study I will keep your decision confidential.

If you have questions, please ask me. If you have any questions later, I can be contacted at zavalina@illinois.edu (email) or 217-265-5406. You may also contact Professor Carole Palmer (email: clpalmer@illinois.edu, voice: 217-244-0653). You may contact the University of Illinois Institutional Review Board (IRB) office (email: irb@illinois.edu, voice: 217-333-2670) for information about your rights in University of Illinois approved research.

You are making a decision whether or not to volunteer. Your signature indicates that you have read and understood the information provided above and have decided to participate. You may withdraw at any time after signing this form. You may keep the attached participant's copy of the form.

I give permission for my interview to be audiotaped ____ (please check to grant consent).

I give permission for my name to be used in connection to this interview/observation ____ (please check to grant consent)

_____

Signature of Participant Date

Appendix C. Coding Manual Used in Transaction Log Analysis of User Queries

A. List of categories to apply:

1. Work
2. Individual person
3. Family
4. Ethnic Group
5. Class of Persons
6. Corporate Body
7. Concept
8. Object
9. Event
10. Place
11. Unknown

B. Coding directions:

Use the following FRBR (1997, 2008) and FRAD (2007) definitions and examples as guidelines for distinguishing between the user search categories:

work — "a distinct intellectual or artistic creation" (FRBR, p. 16) [However, remember that in classification of the collection-level queries, *work* category is broader than FRBR *work* and covers any intellectual or artistic creation that has a title attribute, including the digital collections that are members of the Registry].

individual person — "an individual, encompasses individuals that are deceased as well as those that are living" (FRBR, p. 23), "includes personas established or adopted by an individual through the use of more than one name (e.g., the individual's real name and/or one or more pseudonyms), Includes personas established or adopted jointly by two or more individuals (e.g., Ellery Queen — joint pseudonym of Frederic Dannay and Manfred B. Lee), Includes *personas* established or adopted by a group (e.g., Betty Crocker) (FRAD, p. 13).

family — "Two or more persons related by birth, marriage, adoption, or similar legal status, or otherwise present themselves as a family. Includes royal families, dynasties, houses of nobility, etc. Includes patriarchies and matriarchies. Includes groups of individuals sharing a common ancestral lineage. Includes family units (parents, children, grand children, etc.). Includes the successive holders of a title in a house of nobility, viewed collectively (e.g., Dukes of Norfolk). (FRAD, p. 13)

***corporate body* — "**an organization or group of individuals and/or organizations acting as a unit, encompasses organizations and groups of individuals and/or organizations that are identified by a particular name…" (FRBR, p. 24)

concept — "an abstract notion or idea, encompasses a comprehensive range of abstractions that may be the subject of a *work*: fields of knowledge, disciplines, schools of thought (philosophies, religions, political ideologies, etc.), theories, processes, techniques, practices, etc. A *concept* may be broad in nature or narrowly defined and precise" (FRBR, p.25)

object — "a material thing, encompasses a comprehensive range of material things that may be the subject of a *work*: animate and inanimate objects occurring in nature, fixed, movable, and moving objects that are the product of human creation, objects that no longer exist" (FRBR, p.26)

event — "an action or occurrence, encompasses a comprehensive range of actions and occurrences that may be the subject of a work: historical events, epochs, periods of time, etc." (FRBR, p. 27)

place — "a location, encompasses a comprehensive range of locations: terrestrial and extra-terrestrial, historical and contemporary, geographic features and geo-political jurisdictions" (FRBR, p. 27).

For supersets of individual persons other than ***family*** **and** ***corporate body***, use the following codes:

- ***ethnic group*** (e.g., "Irish Americans", "Sioux Indian", "Basque"),
- ***class of persons*** (e.g., "children that are abused", "prisoners", "country people").

Code fictitious characters on the basis of "what they would be if they really existed" (e.g., Don Quixote would be an ***individual person***, TV series' character Alf, on the other hand, is a creature, just as a dog or a squid, thus a FRBR ***object***).

Code institutions that are not qualified by specific names and locations (e.g., "library", "archive", "can company", "prison") as ***concepts***, code more specifically named institutions (e.g., "Icy Hot Bottle Co.", "library Moorhead") as ***corporate bodies*** or ***objects*** respectively.

For non-English seaarch terms (e.g., French, German, Spanish, Italian, etc.), use a dictionary to find a meaning and categorize accordingly. If it is impossible to categorize a non-English term properly, assign to ***unknown*** category.

In transaction log analysis, code user queries entirely ambiguous as to which search category they belong to (e.g., "aF", "beyond", "LU65") or the intent of the search (e.g., "google", "GEM") into the ***unknown*** search category.

If the user query does not fit into any of the categories listed above but can be meaningfully categorized, create a new appropriate category.

If the user query cannot be assigned to only one of the categories, assign to multiple categories (all that apply).

Appendix D. Coding Manual Used in Content Analysis of Free-text *Description* Collection Metadata Fields

Collection properties coding exercise
April 14, 2010, CIRSS Student Research Group meeting

Please follow the coding guidelines in the end of this document to code the collection-level free-text *Description* fields from American Memory (AM), The European Library (EL), and Opening History (OH) aggregations.

Collection record 1. A Civil War Soldier in the Wild Cat Regiment: Selections from the Tilton C. Reynolds Papers (AM)

Description field:	***Collection properties:***	
A Civil War Soldier in the Wild Cat Regiment: Selections from the Tilton C. Reynolds Papers documents the Civil War experience of Captain Tilton C. Reynolds, a member of the 105th Regiment of Pennsylvania Volunteers. Comprising 164 library items, or 359 digital images, this online presentation includes correspondence, photographs, and other materials dating between 1861 and 1865. The letters feature details of the regiment's movements, accounts of military engagements, and descriptions of the daily life of soldiers and their views of the war. Forty-six of the letters are also made available in transcription.	*1. Audience/uses* *2. Collection development policy* *3. Comprehensiveness* *4. Copyright* *5. Creator of items in collection* *a. Corporate* *b. Individual* *6. Frequency of additions* *7. Funding sources* *8. Geo. coverage* *9. Hosting/contributing institution* *10. Importance* *11. Language of items* *12. Navigation and functionality* *13. Object types/genres* *14. Provenance* *15. Size* *16. Subjects* *17. Temp. coverage* *18. Title* ***19.** Uniqueness* ***OTHER?***	*Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N*

Collection record 2. *Gallica - The digital library of the national library of France (EL)*

Description field:	***Collection properties:***	
Gallica is the digital library of the Bibliothèque Nationale de France (BnF), open to the general public around the world since 1997. It serves as a digital encyclopedia and consists of printed materials (books, journals, newspapers, printed music, and other documents), graphic materials (engravings, maps, photographs, and others), and sound recordings. Gallica makes it possible to find sources that are rare, unusual, out-of-print, or difficult, if not impossible, to access. These materials are royalty-free and available free of charge when used strictly for private purposes. Today, this digital library includes more than 70,000 volumes of digitized texts, 80,000 still images, and 30 hours of sound recordings.	*1. Audience/uses* *2. Collection development policy* *3. Comprehensiveness* *4. Copyright* *5. Creator of items in collection* *a. Corporate* *b. Individual* *6. Frequency of additions* *7. Funding sources* *8. Geo. coverage* *9. Hosting/contributing institution* *10. Importance* *11. Language of items* *12. Navigation and functionality* *13. Object types/genres* *14. Provenance* *15. Size* *16. Subjects* *17. Temp. coverage* *18. Title* ***19.*** *Uniqueness* ***OTHER?***	*Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N*

Collection record 3. *New York State Historical Literature Collection (OH)*

Description field:	***Collection properties:***	
The Cornell University Library New York State Historical Literature is a collection of selected monographs, pamphlets and other materials with expired copyrights chosen from the Cornell Library's extensive collection of New York State Literature. These were materials that were brittle and decaying and in need of rescue. These materials were digitally scanned and facsimile editions on acid free paper were created.	*1. Audience/uses* *2. Collection development policy* *3. Comprehensiveness* *4. Copyright* *5. Creator of items in collection* *a. Corporate* *b. Individual* *6. Frequency of additions* *7. Funding sources* *8. Geo. coverage* *9. Hosting/contributing institution* *10. Importance* *11. Language of items* *12. Navigation and functionality* *13. Object types/genres* *14. Provenance* *15. Size* *16. Subjects* *17. Temp. coverage* *18. Title* ***19.*** *Uniqueness* ***OTHER?***	*Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N*

Collection record 4. *Born in Slavery (AM)*

Description field:	***Collection properties:***	
Born in Slavery: Slave Narratives from the Federal Writers' Project, 1936-1938 contains more than 2,300 first-person accounts of slavery and 500 black-and-white photographs of former slaves. These narratives were collected in the 1930s as part of the Federal Writers' Project of the Works Progress Administration (WPA) and assembled and microfilmed in 1941 as the seventeen-volume *Slave Narratives: A Folk History of Slavery in the United States from Interviews with Former Slaves*. This online collection is a joint presentation of the Manuscript and Prints and Photographs Divisions of the Library of Congress and includes more than 200 photographs from the Prints and Photographs Division that are now made available to the public for the first time. *Born in Slavery* was made possible by a major gift from the Citigroup Foundation.	1. *Audience/uses* 2. *Collection development policy* 3. *Comprehensiveness* 4. *Copyright* 5. *Creator of items in collection* a. *Corporate* b. *Individual* 6. *Frequency of additions* 7. *Funding sources* 8. *Geo. coverage* 9. *Hosting/contributing institution* 10. *Importance* 11. *Language of items* 12. *Navigation and functionality* 13. *Object types/genres* 14. *Provenance* 15. *Size* 16. *Subjects* 17. *Temp. coverage* 18. *Title* **19.** *Uniqueness* ***OTHER?***	*Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N*

Collection record 5. *The Quixote in the National Library of Spain (EL)*

Description field:	***Collection properties:***	
Digitized collection in which you can see the main editions of El Quijote de Miguel de Cervantes. It covers from the first one, done in 1605 by Juan de la Cuesta, to the XIXth century. It includes, among others, the editions of Joaquín de Ibarra, Gabriel de Sancha or the Royal Press, in 1819.	*1. Audience/uses* *2. Collection development policy* *3. Comprehensiveness* *4. Copyright* *5. Creator of items in collection* *a. Corporate* *b. Individual* *6. Frequency of additions* *7. Funding sources* *8. Geo. coverage* *9. Hosting/contributing institution* *10. Importance* *11. Language of items* *12. Navigation and functionality* *13. Object types/genres* *14. Provenance* *15. Size* *16. Subjects* *17. Temp. coverage* *18. Title* ***19.*** *Uniqueness* ***OTHER?***	*Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N*

Collection record 6. *Oliver Family Photograph Collection (OH)*

Description field:	***Collection properties:***	
The Oliver Collection consists of approximately 2700 glass plate negatives and photographic prints taken mainly by amateur photographer William Letts Oliver and his son Roland L. Oliver. The photographs date from the late 1800s to the early 1900s. Subjects include maritime and yachting scenes, views of California and San Francisco Bay area, University of California at Berkeley, mining, logging, the Bohemian Grove, and the Oliver family. Also of note are earlier photographs of Chile and Peru dating from 1860 to 1867. Photographs are attributed to William Letts Oliver, his son Roland Letts Oliver, and occasionally other members of the Oliver family. Other notable photographers represented in the collection include W. H. Lowden and Gabriel Moulin. Also present are professionally produced lantern slides of images by various photographers, and a magic lantern projector and small number of magic lantern slides with illustrations from children's stories**.**	*1. Audience/uses* *2. Collection development policy* *3. Comprehensiveness* *4. Copyright* *5. Creator of items in collection* *a. Corporate* *b. Individual* *6. Frequency of additions* *7. Funding sources* *8. Geo. coverage* *9. Hosting/contributing institution* *10. Importance* *11. Language of items* *12. Navigation and functionality* *13. Object types/genres* *14. Provenance* *15. Size* *16. Subjects* *17. Temp. coverage* *18. Title* ***19.*** *Uniqueness* ***OTHER?***	*Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N* *Y N*

CODES TO USE AND EXAMPLES OF CODING

1. **Audience/use** (e.g., "Alabama residents and students, researchers, and the general public in other states and countries", "for personal use or educational presentations")

2. **Collection development policy** (e.g., "titles published between 1850 and 1950 were selected", "The main geographic focus of the collection is on the region of the state north of Clark County")

3. **Comprehensiveness** (e.g., "a comprehensive and integrated collection of sources on the history and topography of London", "one of the most ambitious and comprehensive effort to date to deliver educational content on the Civil Rights Movement")

4. **Copyright** (e.g., "restricted to items that are not covered by copyright protection")

5. **Creator of items in collection:**

    a. **Corporate** (e.g., "Museum Extension Projects of Pennsylvania, New Jersey, Connecticut, Illinois, and Kansas crafted most of the items")

    b. **Individual** (e.g., "Mr. Cushman extensively documented the United States as well as other countries")

6. **Frequency of additions** (e.g., "Regular additions to the collection are expected", "future filming of some 10,000 volumes per year")

7. **Funding sources** ("digitized as the result of an Illinois State Library FY98 Educate and Automate grant")

8. **Geographic coverage** (e.g., "Mayan city of Uxmal")

9. **Hosting/contributing/participating institution** (e.g., "project brings Tufts, and the Virginia Center for Digital History together with the University to build a digital repository")

10. **Importance** (e.g., "materials are significant in their place within the fabric of American culture", "an archive of unparalleled importance")

11. **Language of items** (e.g., "university listing of faculty and students entirely printed in Latin")

12. **Navigation and functionality** (e.g., "arranged chronologically by Japanese periods", "accessible by date of issue or by keyword searching")

13. **Object types/genres** (e.g., "pamphlets, leaflets, and brochures", "songbooks", "lanterns, torches, banners")

14. **Provenance** ("a 1988 bequest from the collection of Los Angeles architect Rudolf L. Baumfeld", "collection comprises books selected from the Library of Congress's General Collections and from its Rare Book and Special Collections Division")

15. **Size** (e.g., “over 1500 newspaper articles”, “hundreds of personal letters, diaries, photos, and maps”)

16. **Subjects** (e.g., “cover a broad range of topics, including ranching, mining, land grants…”, “as a member of the U. S. Army, 252nd Field Artillery Battalion, he captured images of life as a soldier”)

17. **Temporal coverage** (e.g., “California Golden Rush”, “dating to 1850s”)

18. **Title** (e.g., “The ’League of Nations Statistical and Disarmament Documents’ collection contains the full text of 260 documents…”)

19. **Uniqueness** (e.g., “a number of absolutely unique printed books”, “rare historic published monographs and serials”).

Appendix E. Content Analysis of Free-text *Description* Collection Metadata Fields: Intercoder Reliability Matrix

Collection property	coder agreement (%), record 1	coder agreement (%), record 2	coder agreement (%), record 3	coder agreement (%), record 4	coder agreement (%), record 5	coder agreement (%), record 6	ave. coder agreement (%) for collection property
Copyright	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	**100.00%**
Frequency of additions	100.00%	100.00%	100.00%	87.50%	100.00%	100.00%	**98.00%**
Funding sources	100.00%	87.50%	100.00%	100.00%	100.00%	100.00%	**98.00%**
Language	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	**98.00%**
Navigation and functionality	100.00%	100.00%	87.50%	100.00%	100.00%	100.00%	**98.00%**
Size	100.00%	100.00%	100.00%	100.00%	100.00%	75.00%	**96.00%**
Hosting/contributing institution	100.00%	100.00%	87.50%	100.00%	75.00%	100.00%	**94.00%**
Objects/genres	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	**94.00%**
Temp. coverage	100.00%	100.00%	75.00%	87.50%	87.50%	100.00%	**92.00%**
Audience	100.00%	100.00%	100.00%	62.50%	87.50%	87.50%	**90.00%**
Title	100.00%	100.00%	100.00%	100.00%	50.00%	87.50%	**90.00%**
Creator of items	100.00%	75.00%	75.00%	87.50%	87.50%	100.00%	**88.00%**
Uniqueness	87.50%	87.50%	100.00%	62.50%	87.50%	100.00%	**88.00%**
Subjects	87.50%	87.50%	50.00%	87.50%	75.00%	100.00%	**85.00%**
Collection development policy	87.50%	87.50%	87.50%	100.00%	75.00%	62.50%	**83.00%**
Importance	87.50%	75.00%	100.00%	62.50%	87.50%	75.00%	**81.00%**
Comprehensiveness	62.50%	75.00%	87.50%	100.00%	62.50%	100.00%	**81.00%**
Geo. coverage	50.00%	100.00%	75.00%	50.00%	100.00%	100.00%	**79.00%**
Provenance	87.50%	75.00%	75.00%	87.50%	87.50%	87.50%	**79.00%**
OVERALL INTERCODER RELIABILITY							**90.00%**

Appendix F. Glossary of Important Terms

Collection-level metadata — a structured, open, standardized and machine-readable form of metadata providing a high-level description of an aggregation of individual items.

DCCAP — Dublin Core Collection Description Application Profile, based on the Dublin Core metadata standard, and used for describing digital collections.

Digital collection — in the context of this research, a set of digital items created according to the some collection development criteria, united by the thematic cohesiveness (e.g., by topic area, holding institution, type of materials), searchability as a distinct collection, and a unique point of entry (URL), and included in one of the aggregation of digital collections.

Digital library — the Digital Library Federation (DLF) defines digital library as organization that provides the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community.

Domain knowledge — a searcher's level of knowledge both on a specific search topic and the broader subject domain, domain knowledge influences information seeking behavior and the outcomes of the search for information.

Formalized metadata — metadata elements, the values of which are usually expressed through controlled-vocabulary terms (e.g., "subject" [LCSH, AAT], "geographic coverage" [TGM] etc.)

FRAD — *Functional Requirements to Authority Data* conceptual model (2007)

FRBR — *Functional Requirements for Bibliographic Records* conceptual model (1997, 2008)

Free-text metadata — metadata elements, the values of which are expressed with the natural language, without the restriction to controlled-vocabulary terms (e.g., "description", "title", "notes," etc.)

FRSAR — *Functional Requirements for Subject Authority Records* (2008)

Metadata — a structured data about an object that supports functions associated with the designated object.

Metadata Elements — properties of the object/collection that are defined in a specification. "Author/creator", "title," and "subject" are properties that are commonly identified as metadata elements.

Metadata Record — an organized collection of metadata elements with content values that represent an object or collection (e.g., bibliographic or catalog records, finding aids)

Query — a measure in **Transaction Log Analysis**, string of terms submitted by a user in a given instance of interaction with the system.

Query complexity — a measure in **Transaction Log Analysis**, examines the query syntax, including the use of advanced searching techniques such as Boolean operators, phrased searching, stemming and search limits.

Query frequency — a measure in **Transaction Log Analysis**, number of times query used in a log.

Query length — a measure in **Transaction Log Analysis**, number of words in query.

Richness of collection-level metadata — expression of digital collection's subject matter through mutually complementary values encoded in a variety of collection metadata fields and .representing a variety of collection characteristics in the free-text *Description* fields.

Search term — a series of characters within a query separated by white space or other separator.

Session length — a measure in **Transaction Log Analysis**, the number of queries submitted by a searcher during a defined period of interaction with the system.

Subject access — systematic (e.g., classification system), topical (e.g., subject headings), and natural (e.g., title, abstract words) approaches to the subject matter in a collection, encompasses both processes of subject cataloging and retrieval by the searcher.

Subject entities — entities of the FRBR entity-relationship conceptual model of bibliographic universe (and updated by FRAD model), each of which can be a subject of a work: work, expression, manifestation, item, person, family, corporate body, concept, object, event, and place.

Subject search — a search, where the user seeks to identify resources on a known topic.

Transaction log — an electronic record of the interactions between a system and the users of that system (e.g., between a web site of the digital collection and users searching or browsing for information on that website).

User interactions — the physical expressions of communication exchanges between the searcher and the system (usually recorded in transaction log).

Transaction log analysis — the use of data collected in **transaction log** to investigate particular research questions concerning the interactions of a user with a system.

Unique query — a measure in **Transaction Log Analysis**, a query that is different from all other queries in the transaction log, regardless of the searcher, all identical queries are usually collapsed together to give the unique queries.